

A global folly

If not a global non-nuclear proliferation regime based on international treaties, then what?

North Korea's nuclear test this week (see page 610) is alarming for two reasons: the volatile nature of the nation that has just barged its way into the nuclear-weapons club, and the accompanying sense that the international anti-proliferation regime is slowly unravelling before our eyes.

There is precious little that can be done about the former. However, the enfeebled state of the anti-proliferation regime — ultimately, an even greater threat than the existence of a North Korean weapon — is a problem that can at least be addressed, if the will to do so exists.

It must be fervently hoped that the challenge posed by North Korea will provide the political impetus needed to strengthen anti-proliferation agreements. Further erosion would leave us with a nuclear free-for-all and, ultimately, facing the kind of grim apocalypse seldom contemplated since the end of the cold war.

International treaties concerning nuclear weapons are not particularly sexy or exotic; they are complex and arcane. In our 24-hour news cycle, their advocates are liable to be drowned out by those with shriller voices. That has already happened to a troubling degree in the United States. It is so much easier to call for the bombing of Iranian bunkers, for example, than to argue for the Comprehensive Nuclear Test Ban Treaty (CTBT) to be brought back before the US Senate for ratification.

What good would this piece of paper do, the detractors cry, when faced with the ambitions of North Korea or Iran? It would be fanciful to claim that the existence of any non-proliferation treaty would have halted the North Korean test. However, even a regime such as Kim Jong-il's is not entirely beyond the reach of such documents — North Korea withdrew from the nuclear Non-Proliferation Treaty (NPT) in 2003, before conducting its test.

The best that a set of nuclear-weapons treaties could do in the face of North Korea's ardour is provide a moral and legal framework that allows the rest of the world to pinpoint, criticize and punish the tester with unanimity and conviction. But such a framework would not prevent a test, any more than the enactment of common law prevents every murder.

The 1970 non-proliferation treaty — the main foundation of the tattered anti-proliferation regime — was constructed with the direct involvement of weapons scientists in the United States and the former Soviet Union. These individuals saw clearly the unique moral and military danger of nuclear weapons and, noting the lack of the necessary technical knowledge in diplomatic or other official circles, took it upon themselves to confront it.

Even today, expertise in disarmament and anti-proliferation issues lies chiefly with scientists working at the nuclear-weapons laboratories, with think-tanks or in government. They have looked on aghast as disarmament sceptics around the world have scorned the value of international treaties, rubbished the CTBT on specious technical grounds, and even blamed the alleged toothlessness of the NPT for the gradual spread of nuclear weapons to India, Pakistan and now North Korea. Just as worryingly, the issue of what to do with the arsenals of the existing nuclear-weapons states has slipped off the political agenda.

This week's events are the culmination of an appalling decade that has seen India and then Pakistan carry out nuclear tests with scant punishment. The NPT has languished in the absence of agreement on how to modernize it (see *Nature* 435, 132–133; 2005), and the CTBT has been sidelined. These developments serve to undermine the already faint possibility of a coordinated response to the North Korean test by the United States, China (the two big players in this case) and the rest of the world.

It is imperative that scientists and others with relevant knowledge of nuclear weapons redouble their efforts, and rally to repair years of vandalism inflicted on international anti-proliferation efforts. The goal is to build a treaty structure that will help contain the spread of nuclear weapons, eventually laying the groundwork for their international control and subsequent elimination. That must be the way forward — the alternative is despair. ■

"It is imperative that scientists rally to repair years of vandalism inflicted on international anti-proliferation efforts."

Forgotten plights

Scientists' human-rights groups deserve stronger backing.

"First they came for the Socialists, and I didn't speak up, because I wasn't a Socialist... Then they came for the Jews, and I didn't speak up, because I wasn't a Jew. Then they came for me, and there was no one left to speak up for me." Martin Niemöller's poem, criticizing the inaction of German intellectuals in the face of the rise of the Nazis, serves as a powerful analogy for

why scientists should be concerned by abuses of academic freedom, wherever they occur.

Most readers of *Nature* take it for granted that they can travel to work each day, free to enquire, express opinions and criticize government policy, without fear of intimidation or reprisals — let alone imprisonment or torture. Sadly, these freedoms can only be dreamt of in many countries of the world, where academics must live with, and often suffer directly, human-rights abuses. Their plight is our business.

But beyond humanitarian grounds, in this interconnected world we are engaged in a battle of ideas, and the failure to defend any abuse of academic freedom undermines the very principles that guarantee



the rights we currently enjoy. Oppressive regimes typically stifle enquiry, as critical minds will inevitably also scrutinize their leaders. Enquiry is further undermined in such environments by the award of senior academic posts to the politically loyal rather than the competent, and the selection of policies or actions that suit governments' agendas, regardless of the scientific evidence.

That latter characteristic is central to the trial of six medical workers — five Bulgarian nurses and a Palestinian doctor — currently facing the death penalty in Libya on charges of infecting hundreds of children with HIV (see page 612). The real evidence has been purged from the trial. So it is encouraging that several major scientific bodies have now weighed in to demand that the court hears the scientific facts.

Tripoli may seem far away, but knowledge and academic freedom are central planks in many other struggles across the world for more open, democratic societies. Academics and universities are often hotbeds of such reform movements, and every year hundreds of academics worldwide consequently face threats, or worse. It is important that we do not forget them.

Many learned societies, including the American Physical Society

and the American Chemical Society, as well as several scientific academies, have human-rights committees that play an active role in defending individuals at risk. This diverse range, and the mechanism whereby one body takes the lead on a case where it knows the community, is an effective way of dividing up resources. Cases are many, and no one community can give sustained attention to them all.

Most societies' human-rights activities are run on a shoestring by volunteers. The US National Academies' Committee on Human Rights is among the most effective, and has a full-time secretariat. Yet it runs on a budget of just \$0.5 million a year, most of it contributed by philanthropies. Scientists must find the means to better fund and professionalize such activities.

Often these committees use political contacts and letter-writing campaigns to try to influence the outcome of particular cases. At the very least, this serves to remind perpetrators that they are under international scrutiny. Scientists who have been freed testify that, although difficult to pin down, such support is crucial. All scientists can contribute, by making themselves aware of current cases of human-rights abuses and by lending their support to campaigns against them. ■

Ambassador for Earth

Is it time for SETI to reach out to the stars?

One of the strengths of the community involved in the search for extraterrestrial intelligence — known as SETI — is its imaginative capacity to take seriously things that most people dismiss out of hand.

The idea that the technologies of astronomy might go beyond allowing us new insights into the natural world, and provide us with a means of communion with alien creatures, is not entirely fanciful. Artificial phenomena, such as civilizations, will radiate energy as surely as natural ones do — and may even do so with the intention of communicating. We on Earth send a mish-mash of unnatural-looking radio waves out into the cosmos, not to mention a handful of neutrino beams.

But what if we were to add to that mish-mash some deliberate signals? If beamed tightly enough with the help of a radio-telescope antenna, even a low-power radio transmission can stand out from the general murmur to a star where it is aimed. This is the idea behind so-called 'active SETI', which some enthusiasts think is a way forward for the field. There are others who think that it poses small but real dangers, and thus needs to be discussed more broadly. But at a recent meeting of the International Academy of Astronautics SETI study group in Valencia, Spain, the mood of the gathering was with the enthusiasts.

It is easy to see the appeal of active SETI — it's right there in the adjective. Traditional SETI involves looking at vast amounts of radio data and finding nothing. Active SETI allows you to compose messages, pick target stars, develop new encodings, and so on. It can be used as an outreach tool — the European television channel Arte is currently encouraging people to send it messages specifically to be beamed to the stars as part of the celebrations surrounding the

launch of Corot, a French satellite designed to detect planets around distant stars.

At the same time, though, the risk posed by active SETI is real. It is not obvious that all extraterrestrial civilizations will be benign — or that contact with even a benign one would not have serious repercussions for people here on Earth. There is already an agreement within the SETI community that, should a signal from beyond be picked up, various bodies will discuss what response, if any, should be sent. Yet the Valencia meeting voted against trying to set up any processes for deliberating over the style or content of any spontaneous outgoing messages. In effect, anyone with a big enough dish can appoint themselves ambassador for Earth.

The chances of active SETI causing unpleasant outcomes with today's technology are in fact remote, as this would require us to lift ourselves over the threshold of detectability for an alien civilization that just happened to be orbiting the star at which the message was aimed, or to reveal some peculiar flaw in our psychological make-up that alien 'black-ops' specialists might start working out ways to exploit. Either way, the harm, even if done at the speed of light, would take decades to arrive.

These small risks should nonetheless be taken seriously. When technologies offer radical new possibilities, the people who have the privilege of playing with them also have a duty to consult widely about what those possibilities might mean. The SETI community should assess them in a discussion that is open and transparent enough for outsiders to listen to and, if so moved, to actively participate. Of course, consensus may not always be possible — but the sort of debate out of which consensus has a chance to emerge must now take place. ■

"It is not obvious that all extraterrestrial civilizations will be benign, or that contact with even a benign one would not have serious repercussions."

RESEARCH HIGHLIGHTS

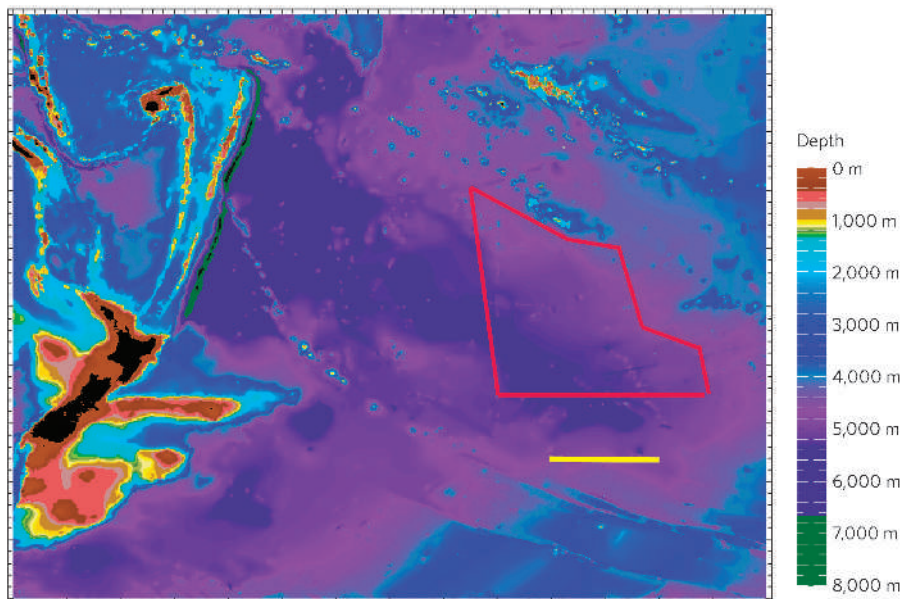
Laid bare

Geology **34**, 873–876 (2006)

In the central South Pacific Ocean lies a swath of sea floor as big as western Europe that is bare of sediment, researchers report.

David Rea of the University of Michigan and his colleagues made the surprising discovery after taking seismic profiles of a little-studied area of sea floor during a research cruise in 2005. The bare zone (indicated by red lines) apparently experiences all the factors — including low biological productivity and a lack of dust input — that prevent sediment from piling up.

Rocks in the bare zone may interest marine geologists studying how seawater interacts with ocean crust. But for those who made the discovery, the absence of sediment is ironic: the team was scouting regions where they might collect sediment cores to learn about past climate. This effort is part of the Integrated Ocean Drilling Program.



REA ET AL./NOAA

VIROLOGY

Crossbred virus

Science **314**, 95 (2006)

A hybrid virus that is part HIV and part SIV has been created by researchers in the United States, and may lead to better animal models of the human disease.

HIV is picky and only replicates in human cells, so AIDS researchers working with non-human primates have had to use SIV, the monkey equivalent. This does not fully mimic the human disease.

Paul Bieniasz at the Aaron Diamond AIDS Research Center in New York and his colleagues replaced two proteins in HIV — one that encapsulates the viral RNA, another that is essential for replication — and allowed the engineered virus to adapt to macaque cells. The hybrid virus contains a genome that is 88% HIV.

ATMOSPHERIC SCIENCE

Vortex threatens ozone

Geophys. Res. Lett. **33**, L18811 (2006)

Swirling winds above the Arctic forced abnormally high amounts of ozone-destroying nitrous oxides (NO_x) into the stratosphere in early 2006, say Cora Randall of the University of Colorado in Boulder and her colleagues.

Measurements showed that NO_x levels in the stratosphere were unusually high during this period, but that the mechanism that produces NO_x in the upper reaches of the atmosphere was not particularly active. Instead, the team argues, extraordinary

weather was to blame. The polar vortex — a persistent, large-scale cyclone near the Earth's pole — was surprisingly strong, and this drew more NO_x down into the stratosphere.

The implication is that if changes in global climate can affect the strength of the polar vortex, this will change the amount of ozone that is destroyed in the already patchy area over the Arctic.

CELL BIOLOGY

How eggs yolk sperm

Nature Cell Biol. **8**, 1143–1148 (2006)

The eggs of *Caenorhabditis elegans* worms attract sperm with the help of fatty molecules stored in their yolk, say US researchers.

Michael Miller of the University of Birmingham at Alabama and his colleagues showed that eggs needed to contain polyunsaturated fatty acids, or PUFAs, for sperm to crawl towards them.

In mammals, PUFAs are converted into signalling molecules known as eicosanoids, which attract immune cells. The team suggests that, in worms and mammals, PUFAs may have a general role as precursors of signals that summon motile cells to target tissues.

COSMOLOGY

Deep, dark picture

Mon. Not. R. Astron. Soc. **372**, 28–32 (2006)

Scientists who pooled data on two types of cosmic explosion have revealed new constraints on the make-up of the Universe.

In 1998, observations of exploding stars

known as type 1a supernovae revealed that the expansion of the Universe is accelerating, driven by an unknown force termed 'dark energy'. But these supernovae aren't sufficiently bright to shine from the earliest times of the Universe. Astronomers hoping to probe the properties of dark energy further back in time have therefore turned to a brighter class of event, known as gamma-ray bursts.

Building on earlier work that proposed a way to calibrate the data, Claudio Firmani of the Brera Astronomical Observatory in Merate, Italy, and his colleagues combined measurements of 115 supernovae and 19 gamma-ray bursts. Their results suggest that the density of dark energy has remained unchanged over about 10 billion years.

ANALYTICAL CHEMISTRY

Blast it

Angew. Chem. Int. Ed. doi:10.1002/anie.200602449 (2006)

Blasting ionized molecules out of a sample is an effective way of making microscopic compositional maps of biological tissues, according to R. Graham Cooks and his colleagues at Purdue University in Indiana. They have adapted a relatively new method of mass spectrometry, known as desorption electrospray ionization (DESI), to create an imaging technique with a resolution of less than 0.5 millimetres.

DESI uses a jet of charged solvent droplets to expel molecular ions and measure their mass, and thus chemical formulae. The researchers scanned the spray nozzle over

slices of rat brain and characterized the various lipid ions ejected from each pixel area. The resolution is lower than for related ionization imaging methods, but the sample needs no pretreatment and the images are obtained under ambient conditions, making *in vivo* use a possibility.

NANOTECHNOLOGY

Doubled up

Nano Lett. doi:10.1021/nl061898e (2006)

Fibres made with two semiconductors running side-by-side could be efficient devices for degrading organic pollutants, researchers report.

Zhaoyang Liu and Darren Delai Sun of Nanyang Technological University in Singapore and their colleagues used a technique known as electrospinning to make fibres about 100 nanometres wide that contained two thinner threads fused together. One of these was made from titanium dioxide (TiO₂), the other tin dioxide (SnO₂).

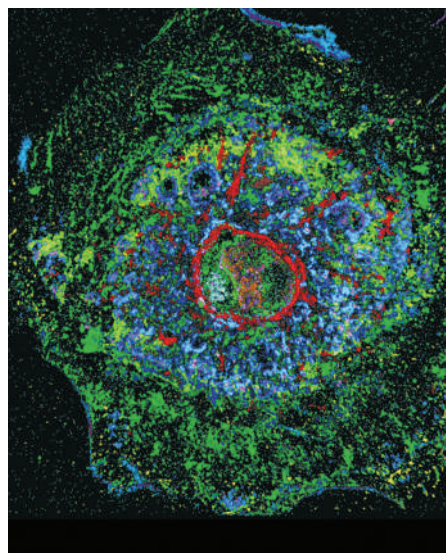
SnO₂ enhances the photocatalytic activity of TiO₂, which oxidizes organic materials using the energy from ultraviolet light. The twin fibre also has a greater surface area on which reactions can occur than other composite structures, such as bilayer films.

MICROBIOLOGY

Unclogging severe malaria

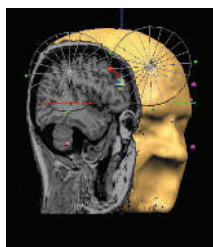
PLoS Pathogens 2, e100 (2006)

Researchers in Sweden may have found a way to eliminate the nasty side effects of the malaria drug heparin. These side effects, which stem from the drug's anticoagulant activity, stopped its use.



Severe malaria occurs when red blood cells infected with the parasite *Plasmodium falciparum* stick to each other and to vascular walls. This causes a host of problems, including anaemia and respiratory difficulties. Heparin disrupts this binding, but can also cause severe bleeding.

A team led by Mats Wahlgren of the Karolinska Institute in Stockholm has now shown that treating heparin with the compound periodate destroys its anticoagulant activity but retains its malaria-fighting properties. In tests on macaques, the modified drug was able to unstick infected cells and restore blood flow.



NEUROSCIENCE

Fair game

Science doi:10.1126/science.1129156 (2006)

Researchers have identified a brain region that has a critical role in our

desire to punish unfair behaviour.

Daria Knoch and Ernst Fehr of the University of Zurich and their colleagues studied people playing the ultimatum game, in which one participant decides how to share a sum of money with a fellow player. If that player feels the offer is unfair, they can prevent the other participant from keeping any money by choosing to forfeit their own share.

When magnetic stimulation was used to inhibit activity in the brain's right dorsolateral prefrontal cortex, players become more likely to accept unfair offers.

PROTEOMICS

Mapping togetherness

Nature Biotechnol. doi:10.1038/nbt1250 (2006)

A method to map a hundred or more proteins in a single cell or tissue slice has been developed by Walter Schubert of the University of Magdeburg, Germany, and his colleagues.

They designed robotic workstations to carry out repeated rounds of antibody labelling and fluorescent imaging, identifying different proteins in sequence. A novel algorithm then analysed the images, pixel-by-pixel, to provide a map of protein networks in the cell or tissue (pictured left). The scientists applied the technique to compare protein distributions in diseased and healthy tissues.

It is hoped that such methods will help to unravel how the hundreds of thousands of proteins in a cell interact with each other, eventually in real time.

C. EISENEGGER

JOURNAL CLUB

Brent R. Stockwell
Columbia University, New York, USA

A chemical biologist is drawn to a new way of validating cancer drug targets.

I am interested in translating basic research into cancer therapies, so I am always looking for new approaches to selecting drug targets. Recently, I read of a technique that promises to help us cope with changing ideas about the types of target we should go after.

Typically, researchers have concentrated on finding proteins that trigger the genesis of tumours; the products of oncogenes. Some subsets of these 'oncoproteins' can be targeted by a class of drug known as small molecules.

My lab's drug-discovery strategy involves the use of synthetic lethal screens. We screen small molecules to identify those that are lethal to tumour cells with a specific oncogene.

I have noticed that, in tumour cells, effective small molecules often target proteins that are not products of oncogenes. These other proteins are not involved in tumorigenesis, but seem to be required for tumour maintenance.

In May, researchers reported a way of testing such 'tumour maintenance' genes in mice (K. Politi *et al. Genes Dev.* 20, 1496–1510; 2006).

They used a switch known as an inducible promoter to turn a classic oncogene on and off. The oncogene, a mutant epidermal growth factor receptor (EGFR), triggers lung cancer by causing excessive cell proliferation. Turning off the gene in existing tumours in mice tested whether the mutant EGFR was also required for the tumour's maintenance. The team found that it was.

Although oncogenes are crucial to tumour formation, it is the tumour maintenance genes that we should target with therapeutic drugs. Here, we have a way to validate such targets that I encourage my colleagues in the translational cancer field to adopt.

SPECIAL REPORT

The fizzle heard around the world

North Korea's nuclear test raises more questions than answers. Despite the small size of the blast, **Jim Giles** and **Geoff Brumfiel** get little reassurance from the weapon watchers.

It was probably the worst first nuclear test in history, say weapons experts who have carried out preliminary analyses of the seismological data. But the North Koreans will have learned a great deal from the nuclear blast they set off on 9 October. And although experts' estimates of the country's technological expertise aren't much altered, North Korea's intent is now unmistakable.

Seismographs around the world registered shock waves from the event, which the North Korean press agency claimed originated from an underground nuclear test (see Map). The event, measured at magnitude 4.2 by the US Geological Survey, was probably too large to be caused by conventional explosives, but too small to be a completely successful nuclear test — unless, as some experts believe, North Korea was testing a more sophisticated device than previously thought.

As *Nature* went to press, there was disagreement over the size of the explosion (see 'The nuclear detectives'), although most governments and experts agree that the seismograph data suggest a device equivalent to around 500 tonnes of TNT. If that is the case, the test was probably what nuclear experts term a 'fizzle' — a device that failed to create a full chain reaction in the plutonium and so produced a smaller than planned nuclear explosion. Weapons experts



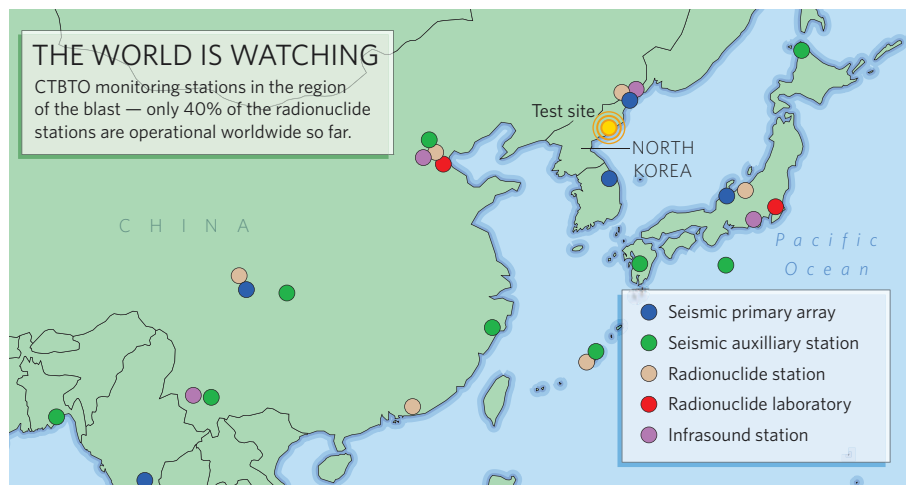
South Koreans express their outrage at North Korea's nuclear test (see seismogram, right).

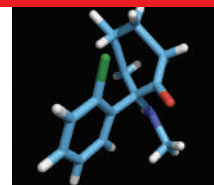
assumed that North Korea had one or two nuclear devices, but the fizzle suggests that their programme might not be as sophisticated as thought. "It looks like it was the least successful first test ever," says James Acton, nuclear-weapons expert at London think-tank VERTIC.

"They're an embarrassment to the nuclear club," adds Jeffrey Lewis, executive director of the Managing the Atom project at Harvard's

Belfer Center for Science and International Affairs. North Korea's bomb is thought to have been a relatively simple atomic device, working entirely on nuclear fission, in which the margin for failure can be reduced by adding extra material and conventional explosives. Lewis says he is unaware of a first test by any other country leading to such a failure. "Frankly I'm baffled," he says. "This is supposed to be really easy."

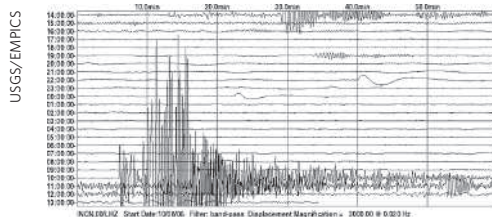
Experts guess that the North Korean device consisted of a series of conventional explosives that were detonated to compress a sphere of plutonium until it reached critical mass, at which point a runaway fission reaction causes a massive release of energy. But even with such a simple device, there can be problems. If more than a few percent of the fissile material consists of plutonium isotopes other than plutonium-239, the chain reaction can begin before the sphere reaches maximum compression. The neutron generator, which triggers the chain reaction, might have been triggered at slightly the wrong time. It's also possible that in an effort to conserve their limited supply of plutonium, the North Koreans put too little in their device. They might also have used too few conventional explosives or these did not work





KETAMINE
From rave drug to antidepressant?
See News Feature on page 629

T. EVANS/SPL



properly, leading to an asymmetrical implosion. "It's not that difficult, but a lot of things can go wrong," says William Press, a senior fellow at Los Alamos National Laboratory (LANL) in New Mexico, and a member of the JASONs, a group of scientists that provides advice to the US government on nuclear issues.

The much more worrying possibility is that North Korea was testing a more sophisticated weapon. John Pike, director of GlobalSecurity.org, a Washington-based security website, believes it may have been trying to test the first stage, or 'primary', of a hydrogen bomb. In an H-bomb, the fission explosion is designed to trigger a much larger fusion reaction. Lewis discounts the idea: "I don't know of a country that's gone straight to the hydrogen bomb," he says. Press adds that even a primary must have a yield higher than that detected on 9 October: "By far the most likely design is an unboosted, spherical-

The nuclear detectives

As *Nature* went to press, national governments had announced a range of figures for the size of the explosion, although most put the device at equivalent to around 500 tonnes of TNT. Nuclear experts who spoke to *Nature* put the figure in the same range. But the Russian government is reported to have given a much higher figure, around 15,000 tonnes of TNT.

One explanation for some of the variation could be uncertainties over the geology of the test site, which seismograph readings identify as in the north-eastern corner of North Korea (see Map).

The equation relating the size of nuclear explosions to seismograph readings requires a parameter that corrects for local geology. Peter Zimmerman, an arms-control expert at Kings College London, says that if no correction is used, a reading of magnitude 4.2 gives a device of 500-1,000 tonnes TNT equivalent. If the test was situated in soft

rock, that figure could rise to 10 kilotonnes. But geologists contacted by *Nature*, including Paul Richards at Columbia University in New York, who is analysing the seismological data, say the Korean peninsula is made up of hard rock that would propagate signals well.

Some of the uncertainty surrounding the test will be cleared up when radionuclides emitted by the explosion are picked up by monitoring stations, which is expected to happen within a few days. As well as its seismological and infrasound stations, the Comprehensive Test-Ban-Treaty Organization (CTBTO) based in Vienna, Austria, is developing a global network of 80 radionuclide-monitoring stations that collect samples by sucking air through a sheet of paper a metre square. The paper is removed each day and tested for distinctive nuclides, such as xenon, produced by nuclear explosions.

The ratios of the different isotopes detected should, in principle, allow governments

to determine the amount of plutonium burned in the test and how successful the test was, and perhaps the kind of design the North Koreans used. But officials at the CTBTO caution that the network of 80 stations is still being set up, and is only 40% complete. Of the six stations around North Korea, only two are operational, both in Japan. Weather reports suggest, however, that the prevailing winds will blow particles towards the north-east, and the nearest station in that direction is in eastern Siberia, nearly 4,000 kilometres away. "This test came a bit too early for us," says a CTBTO official.

The United States will also get its own radionuclide data from aircraft, and the Chinese and Russians may be able to glean information from military monitoring stations close to the border with North Korea. Japan is also flying planes daily to monitor radiation levels in the air, and says it has detected nothing so far.

implosion, plutonium bomb," he says.

But Pike points to evidence suggesting that North Korea might actually have carried out its first nuclear test in Pakistan in 1998, when the two states' nuclear programmes were closely entwined. Moving from a first nuclear test to a primary might not be that difficult, he says — for example, the Israelis started building a hydrogen bomb in the 1980s, after just one nuclear test in 1979.

Siegfried Hecker, former director of LANL, visited nuclear scientists in North Korea in 2004 and 2005. He too warns against underestimating them. "I would have to assume they have the know-how to make a bomb," he says. "We just don't know exactly what they were trying to do." Hecker thinks it's unlikely the North Koreans were testing a primary. But he says the size of the explosion suggests that they may have been trying to miniaturize the bomb into something that could be carried on a ballistic missile. "That ups the difficulty significantly," he says.

Even if the explosion was a partially failed test of a basic device, North Korean researchers will have learnt a great deal from it, says Acton. Comparing the results of the test with previous

computer simulations will allow the North Koreans to assess the accuracy of their models. Even if the test reveals a seriously flawed design, Acton says that at an “absolute upper limit” the country is no more than two years away from developing a functioning weapon. “If it was a fizzle, then they could easily learn enough from it to be in business for their next test,” agrees Press.

According to most estimates, North Korea has enough material for between six and eight nuclear weapons. It may be hoping to mount those weapons on missiles such as the TaePodong-2, although the bomb tested on 9 October was probably far too big. The TaePodong-2 itself failed and fell into the Sea of Japan after just a 42-second flight when tested on 4 July.

Ultimately, the failure or otherwise of the North Korean weapon may have little effect on the response of nations such as the United States and China, according to Michael Levi, a physicist at the independent Council of Foreign Relations in New York: "From a political standpoint it makes essentially no difference whether it was a fizzle," he says. "What matters is their willingness to test." ■

See Editorial, page 605

Protests mount against Libyan trial

Scientific and human-rights organizations have rallied to the cause of six medical workers imprisoned in Libya, accused of deliberately infecting hundreds of children with HIV. The workers look likely to be found guilty and could face the death penalty, yet independent scientists who have reviewed the evidence are convinced they are innocent.

Emmanuel Altit, a lawyer with *Avocats sans Frontières* based in Toulouse, France, which is handling the defence of the health workers, welcomes the renewed pressure as “formidable, exactly the sort of thing that’s needed now”. But he adds that pressure needs to be increased and sustained throughout the rest of the trial.

The five Bulgarian nurses and one Palestinian doctor, who have been in prison since 1999, were sentenced to death in 2004. The verdict was overturned in 2005 by the Libyan Supreme Court, which ordered a retrial that began in May this year. A verdict is expected next month, and the defence team isn’t hopeful.

Although many human-rights groups have supported the six in the past, the Supreme Court ruling led to a false sense of security and international pressure to free the medics dropped, says Zafra Lerman, chair of the subcommittee on scientific freedom and human rights of the American Chemical Society (ACS). The call by a Libyan prosecutor at the end of August for the death penalty came as a shock. “We relaxed too much; we thought the retrial would just see them being released. We didn’t see this new danger coming,” says Lerman, who has won awards for her work in helping to free Soviet dissidents and Chinese scientists jailed after the 1989 Tiananmen Square protests.

So last week, the American Association for the Advancement of Science, the ACS, Physicians for Human Rights, and Amnesty International issued new human-rights action alerts for the medical workers. In particular, they reiterated last month’s call by the defence lawyers for the court to hear independent



What can pressure by the scientific community achieve?

Taye Woldesemayat, an academic who was freed from an Ethiopian jail in 2002 after six years, says he is living proof that campaigning by scientists for colleagues whose human rights have been compromised can be effective.

“In prison, I felt hopeless at the beginning; I knew they were going to kill me,” he told *Nature*. “But then the letter-writing campaigns began, and as the letters started flowing in [to the Ethiopian government] it was fantastic, I knew I could get out.”

Woldesemayat was jailed in Addis Ababa in 1996 on charges of terrorism and armed conspiracy against the state. Human-rights organizations concluded that he had simply called for greater social justice and democracy.

The American Association for the Advancement of

Science (AAAS) and the US National Academy of Sciences weighed in alongside human-rights bodies such as Amnesty International, and the Ethiopian High Court overturned the charges in 2002. After his release, Woldesemayat moved to the United States, although the Ethiopian government is again pursuing him after he helped form an opposition party last December.

Defending the human rights of scientists, engineers and physicians is a little-known sideline of many scientific organizations. Every year they campaign on hundreds of cases.

The biggest players are the AAAS, which launched its Science and Human Rights Program in 1993, and the US National Academies’ Committee on Human Rights (CHR). Also in 1993, the CHR created a

global action network of 70 science academies, although only 40 participate actively. “Nobel prizewinners and scientists, especially when they represent many different countries, have the power to influence,” says Dagfinn Føllesdal, a philosopher at Stanford University in California and a member of the network’s board.

Carol Corillon, director of the CHR, says science bodies vet cases carefully before taking them on. “If Amnesty International has adopted the case I feel fairly comfortable, as they do good research,” she says. “But we always try to get verification with the families, the lawyers, and people we know in the country.”

Once a case is adopted, lobbying is mainly through private political contacts and

letter-writing to politicians. The impact of such campaigns is hard to measure, admits Corillon. “About three-quarters of our cases get resolved, but it’s difficult to know what role, if any, our action played.”

Corillon feels that scientific academies are often better placed to act in private. But, she adds, “Often a carrot-and-stick approach is most effective. I think that without having the public pressure as well on many cases, we wouldn’t have had success.”

Even when campaigns do not resolve a case, they are crucial, says Claude Cohen-Tannoudji, a 1997 Nobel laureate in physics and board member of the academy network. “It’s important that when someone is persecuted, the government doing the persecution knows that the world is watching.” D.B.

M. TURKIA/AFP/GETTY



Three of the prisoners on trial: Nasia Nenova (right), Snezana Dimitrova and Ashraf Hajjaj in court last month.

scientific evidence (see *Nature* **443**, 245–246, 254; 2006).

But despite a Palestinian being one of the accused, science bodies in the Arab world have been relatively silent on the case. “The case is not on the radar at all,” says Moneef Al-Zou’bi, director-general of the Islamic Academy of Sciences, based in Amman, Jordan. “It has not attracted as much attention in this part of the world compared to elsewhere.”

The academy, created by the Organization of the Islamic Conference (OIC), which represents 57 Islamic countries, has not taken a position on the case. “People here have so many political and economic problems, including the conflicts in Iraq and Palestine, that the case has just not become a priority,” says Al-Zou’bi. Such difficulties mean that the academy’s human-rights focus is much more on promoting basic socio-economic rights, he adds, such as access to clean water.

Not everyone agrees with that position. “This is a difficult call,” says one official at an international human-rights body, who asked not to be named. “Of course it is totally legitimate to prioritize the most basic rights for the large-

est number of people. But there is definitely a pattern of organizations in this region virtually ignoring violations of civil and political rights in their own neighbourhoods. I would think that at the very least, scientific or medical organizations would come out in support of colleagues sentenced to death in such a blatant case of scapegoating and dismissal of science. And in the end, when this can happen to health workers, the right to health also comes under threat.”

The OIC’s Commission on Scientific and Technological Cooperation (COMSTECH) hasn’t taken a position either. Chairman Attaur-Rahman, who is also Pakistan’s higher education minister, says that human-rights issues are beyond the commission’s remit. “COMSTECH’s charter is confined purely to matters concerning scientific and technological cooperation between member states”, and so does not extend to legal and judicial matters, he explains.

Imad Khatib, secretary-general of the Palestine Academy for Science and Technology in East Jerusalem, says his academy “cares deeply about the plight of the nurses and the doctor and that they have a fair trial”, although he says that most of its human-rights efforts are spent on the abuse of scientists’ rights within Palestine. ■

Declan Butler

See Editorial, page 605.



AMERICAN SOCIETY OF HUMAN GENETICS
Check our blog for diary reports from the meeting in New Orleans this week.
www.nature.com/news

ERNEST N. MORIAL
CONV. CENTER

Nobel prize blurs boundaries

At first glance it seems that this year's Nobel Prize in Chemistry has gone to a biologist. Roger Kornberg at Stanford University, California, won the prize for unravelling the mysteries of transcription — getting information out of DNA and into proteins via RNA.

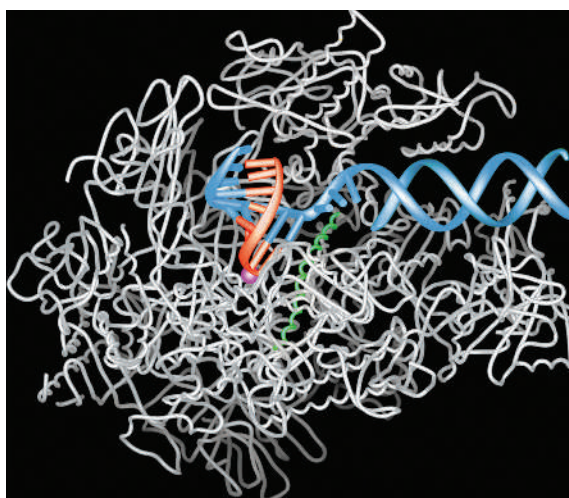
His major breakthrough was working out the crystal structure of RNA polymerase II — an enzyme that helps messenger RNA take information from DNA.

People working in the field are thrilled at the news. "Kornberg's work is a *tour de force* in understanding how transcription works at an atomic level," says Richard Treisman at Cancer Research UK's transcription laboratory in London.

But chemistry bloggers are disgruntled that Kornberg's use of a chemistry tool — crystallography — was enough to warrant a chemistry Nobel prize. Unofficial mutterings from chemistry department corridors confirm some surprise at the choice of recipient. "It is certainly on the biological side of biochemistry," says Malcolm Green, an inorganic chemist from the University of Oxford, UK.

Jesper Svejstrup, a gene-transcription researcher at Cancer Research UK, was a postdoc in Kornberg's laboratory from 1993 to 1996. "It is unusual to have a single winner," he says. "I think the reason is that it's a chemistry prize." Had the prize been for physiology or medicine it would have been shared with Kornberg's co-workers, whose roles were more biological, he suggests.

"What we need is a biology prize," jokes Robert Schrock, an inorganic chemist at Massachusetts Institute of Technology, who won



Prizewinning polymerase: but is it chemistry?

the prize for chemistry in 2005. Schrock's prize, shared with Yves Chauvin and Robert Grubbs, was for hard-core chemistry — metathesis, a catalytic reaction used by organic chemists as a way of swapping different groups of atoms in two reacting molecules. This delighted many chemistry fundamentalists, says Schrock, who has heard plenty of grumbles about biology encroaching on chemistry.

The 2005 prize came after a run of bio-related chemistry prizes: in 2002 for determining the structure of biological macromolecules; in 2003 for deciphering channels in cell membranes; and in 2004 for the discovery of the role of ubiquitin in intracellular protein degradation. "Chemistry is becoming more applied," Schrock suggests as one reason for the trend. And, after all, "biology is chemistry," he insists.

Chemistry is evolving naturally, says Schrock. There are fewer developments in basic chemistry than in its biological applications, which means that fundamental contributions to chemistry get less recognition. It's a matter of labels, he says. "Chemistry is only going to become more important — however you want to name it."

The Nobel committee for chemistry 2006 is chaired by a professor of theoretical physical chemistry, Håkan Wennerström, from Lund University in Sweden. The rest of the members include: a professor of biophysics, a professor of organic chemistry, a professor of molecular biophysics, a professor of physiological chemistry, and a professor of biochemistry. "No surprise that the prize goes to biostuff,"

laments one blogger.

But Aaron Klug, who has worked with Kornberg and who won the chemistry prize in 1982 for developing crystallographic electron microscopy and working out the structure of nucleic acid-protein complexes, is happy with the choice. "Kornberg certainly thinks like a chemist," he says. "The whole way he handled the RNA polymerase material is beautiful chemistry."

Klug goes as far as to suggest that this year's physiology and medicine prize, awarded to Andrew Fire and Craig Mello for discovering RNA interference, could also have been a chemistry prize. "The question is whether large molecules are part of chemistry. And of course they are."

Katharine Sanderson

Intelligent design gets political

After a federal court ruled that intelligent design could not be taught in schools in Dover, Pennsylvania, many thought the idea would fade from public view (see *Nature* 439, 6-7; 2006).

But intelligent design is re-emerging as a political issue. "It continues to have legs, and we're going to have to worry about it," says Lawrence Krauss, a theoretical physicist at Case Western Reserve University in

Cleveland, Ohio, and chair of the advisory board of Help Ohio Public Education (HOPE), a group that opposes the teaching of intelligent design in the state.

In 2004, the Ohio school board approved language encouraging a "critical analysis" of evolution and a lesson plan that included arguments used by supporters of intelligent design. The plan was repealed shortly after the Dover decision in February, but

advocates of intelligent design are again running for positions on the board. In response, HOPE is endorsing its own candidates.

In Michigan, the Republican candidate for state governor, Dick DeVos, has expressed support for teaching intelligent design. He is neck-and-neck with the Democratic governor Jennifer Granholm, and the statement may be an effort to energize conservatives, says Robert Pennock, president of the

pro-evolution Michigan Citizens for Science. "This issue is seen by fundamentalists and evangelicals as part of the culture wars," Pennock says.

But he thinks the endorsement may backfire: "Economic issues are the most important." By endorsing intelligent design, DeVos could lose the backing of Michigan's thriving biomedical industry. Geoff Brumfield

SCORECARD

**Hydrogen**

Hydrogen has been unfairly maligned as the cause of the Hindenburg airship disaster, according to Martin Chávez, mayor of Albuquerque. In the interests of promoting hydrogen fuel, he has called on the US government to pardon the gaseous element.

**Superconductors**

The authors of a citation study that predicted no more papers in high-temperature superconductivity after 2015 seem to have yielded to the wrath of physicists in the field (see *Nature* 443, 376; 2006). They have now removed the extrapolation to zero from their preprint.

**Canada**

It is usually considered that Canada is more environmentally friendly than the United States. But during fishing negotiations at the United Nations last week, Canada became the bad guy. It refused to sign up to a US-backed ban on bottom trawling, which devastates fish populations and the ocean floor.



P. MORRIS/ARDEA.COM

ZOO NEWS

Baby gorillas

Two orphaned gorillas raised by the John Aspinall Foundation (JAF) and released into the wild in the Republic of Congo have both been spotted with young offspring — only the second and third recorded births to reintroduced gorillas. JAF researchers are now planning DNA tests of the males in the group, to identify who the fathers are.

Nano spiders

Readers with arachnophobia may wish to avert their eyes. Researchers have created 'molecular spiders' with legs just 10 nanometres long. Biochemist Milan Stojanovic hopes the spiders could perform certain tasks: "We could have a simple predator-prey system in which one would try to cleave the legs of the other."

Sources: City of Albuquerque, CTV, JAF, BBC

Hard-hitting endeavour captures Ig Nobel

BOSTON

When Ivan Schwab first learned that he was slated to win an Ig Nobel prize for explaining why woodpeckers don't get headaches, his response was perhaps predictable for a man whose past had come back to haunt him: he denied everything.

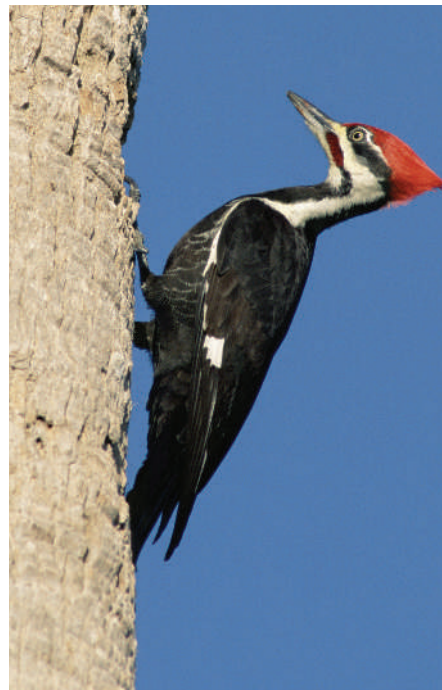
"I didn't do the work," Schwab, an ophthalmologist at the University of California, Davis, told the event's organizer after being informed of the dubious honour in store for him. Ig Nobel prizes are given for "achievements that make people laugh, and then make them think". Schwab gave credit instead to Philip May, a former psychiatry professor at the University of California, Los Angeles, who died in 1986. A deal was soon reached: Schwab agreed to share the 2006 Ornithology prize with May for research initiated by May and associates in 1976 and taken up by Schwab a quarter of a century later.

To May, the woodpecker represented a unique "experiment in Nature" and a chance to understand how an animal could continually use its head as a battering ram without sustaining headaches, concussions or other brain injuries.

A pileated woodpecker may strike its beak against a tree 12,000 times a day for the purposes of feeding, nest construction, ritual drumming (to claim territory and attract mates), and even to relieve tension. They peck at rates of up to 20 times a second — each blow is comparable to striking a wall, face-first, at 25 kilometres an hour. This results in deceleration forces of about 1,200g, which is hundreds of times more than astronauts endure. Given the ill effects of headbanging seen in humans, May wondered why the countryside was not "littered with dazed and dying woodpeckers".

Clearly, woodpeckers do not get headaches or serious head injuries, May reasoned, or they would stop pecking. So how do they prevent it? After dissecting the heads and beaks of two woodpeckers (which were compared with similar toucan samples) and analysing high-speed photographs of live woodpecker strikes, May and his colleagues advanced several explanations (P. R. May *et al. Lancet* **i** (7957), 454–455; 1976; *Arch. Neurol.* **36**, 370–373; 1979).

The woodpecker brain has relatively little cerebrospinal fluid, reducing the transmission of shock waves, and it is tightly packed in



R. NUSSBAUMER/NATUREPL.COM

Striking stat: the pileated woodpecker hits its beak against a tree around 12,000 times a day.

dense, spongy bone that keeps internal movements to a minimum. Muscles encircling the skull serve as shock absorbers, "holding the bill in resilient rigidity".

Like boxers, woodpeckers tense their neck muscles before absorbing a blow. And their drilling trajectory is essentially linear, with little or no rotation of the head, to avoid shearing forces. A third eyelid closes a millisecond before each strike to secure and shield the eye.

Schwab, a bird enthusiast since childhood, read May's 1976 *Lancet* article and later papers with keen interest. On occasion, he mentioned the woodpecker example in ophthalmology classes and at conferences, citing the potential for retinal damage, and always engaged his audience. Yet he yearned to do more on the subject.

He got his chance in 2000, when he became an editor of the *British Journal of Ophthalmology*. In 2002, Schwab put a woodpecker on the journal's cover and wrote an essay presenting May's findings and discussing anatomical features that might provide further protection, as a result of dissections and CT scans he had performed of a pileated woodpecker head

nature podcast



IG NOBEL PODCAST

Listen to interviews with the winners and more...

www.nature.com/nature/podcast

(I. R. Schwab *Br. J. Ophthalmol.* **86**, 843; 2002).

But Schwab's main objective was to bring attention to May's neglected work. "We can't all win Nobel prizes," he says. "But we can look at the world around us and ask questions, just as May did." In his own field, Schwab asks, for example, why polar bears don't get ultraviolet keratopathy — sunburn of the eyes. How can a jellyfish, which lacks a brain, process the visual information it gets from its eyes? Why do scallops have hundreds of eyes and giant clams thousands? Questions such as these, says Schwab, are "the basis of discovery".

For that reason, Schwab says he is proud to be associated with May's research and was delighted to receive an Ig Nobel on his behalf at the ceremony held at Harvard University on 5 October (see 'The Ig Nobel winners' circle'). He certainly got into the spirit of things, attending the festivities in a woodpecker headdress. As for his initial reluctance, Schwab explains: "I was embarrassed to receive an Ig Nobel prize for work I did not do." In that respect, he stands out from previous winners, who were embarrassed to receive Ig Nobels for work they did do. ■

Steve Nadis

The Ig Nobel winners' circle

On 5 October, more than 1,000 people jammed into Harvard's Sanders Theatre to witness the Ig Nobel prize ceremony. Scientists from across the globe gathered to receive awards for offbeat research that would never get through the door at Stockholm. Here are a few winners.

● **Acoustics** Lynn Halpern of Harvard Vanguard Medical Associates and Randolph Blake of Vanderbilt University accepted the award for experiments investigating why people hate the sound of fingernails scraping on a blackboard. Blake speculated that the sound resembles "non-human primate warning calls" and triggers a vestigial fear response. To illustrate the point, the seven Nobel

laureates present reproduced the offending sound on individual blackboards — after donning protective earmuffs.

● **Biology** Bart Knols and Ruurd de Jong of Wageningen Agricultural University in the Netherlands captured their award for showing that female malaria mosquitoes find the smell of Limburger cheese and human feet equally appealing. A synthetic blend of the chemicals involved might help fight malaria by luring the insects away from people. In his acceptance speech, Knols apologized to his wife for forcing her to sit naked in a net with bloodthirsty mosquitoes, as he recorded where they bit.

● **Literature** Princeton psychologist Daniel Oppenheimer triumphed for

a study, published in *Applied Cognitive Psychology*, entitled "Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly". Oppenheimer's acceptance speech: "My research shows that conciseness is interpreted as intelligence. So thank you."

● **Medicine** University of Tennessee's Francis Fesmire won for being, in his words, "the first person to terminate hiccups with digital rectal massage". A trio from Bnai Zion Medical Center in Haifa, Israel, shared the award. At first, Fesmire confessed, he was disappointed to win an Ig Nobel rather than a real one. But his son cheered him up: "It's like winning a Darwin Award, and you don't have to die." **S.N.**

Democrats demand inquiry into climate paper cover-up

Democratic politicians have slammed the US Department of Commerce after officials blocked the release of a statement describing hurricanes' link with global warming.

Fourteen senators have asked the department's inspector-general to launch an investigation into the document, which was to clarify the position of scientists within the National Oceanic and Atmospheric Administration (NOAA). The statement was intended for publication at the start of the hurricane season in June, but was only made public on 3 October after *Nature* revealed commerce officials had blocked its release (see *Nature* 443, 378; 2006).

Bart Gordon (Democrat, Tennessee) of the House science committee also weighed in, asking NOAA on 4 October to provide him with documentation relating to the "suppression" of the fact sheet.

Sanger Institute to shed staff in strategy rethink

More than 60 staff at the Wellcome Trust Sanger Institute in Cambridge, UK, face an uncertain future after being told that their

research is to be phased out. A change in the institute's strategy, announced on 4 October, will result in the culling of five groups that specialize in proteomics and genome biology.

The institute — a centre that played a leading role in sequencing the human genome — will instead focus on studies of natural genetic variation in humans and pathogens, and experimental variation in model organisms such as zebrafish.

Some staff from the five groups may lose their jobs, but Sanger officials say that the institute's overall head count will remain at about 800 over the coming five years, and that they hope to find new positions for most of those affected.

Japan reforms screening to speed up drug approval

Responding to growing criticism, Japan's health ministry plans to overhaul its drug-screening process in order to speed up the delivery of new drugs to market.

On 1 October, the ministry set up a 13-member office to halve the screening period from the current two years to about one year, the same level as the United States. The ministry will reform the methods of screening and the structure of the organization, and create detailed guidelines

so that drug-makers can more efficiently collect the clinical data needed for submission. Currently, manufacturers are often asked to re-file data, causing delays.

The ministry also plans to speed up approval of drugs developed abroad. Despite being the world's second-largest drug market after the United States, 30% of the world's top-selling drugs remain unapproved in Japan.

Anti-global warming ideas enrage French scientists

Climate scientists in France are up in arms over comments made by former French science minister Claude Allègre sceptical of global warming.

On 21 September, in his weekly column in the news magazine *L'Express*, Allègre scoffed at the idea that Mount Kilimanjaro's snowcaps were retreating because of global warming. Allègre went on to question whether



Claude Allègre is at the centre of climate row.

Parks cut profit-share deal with biotech firms

Companies wanting to make money from organisms found in national parks would have to share the wealth with the parks under a new US National Park Service plan.

The draft plan, published late last month, triggered a fresh outcry from not-for-profit groups. Seven years ago, they won a federal lawsuit that forced the park service to carry out an environmental assessment of the effects of such commercial activity in parks. "Commercial bioprospecting undermines the basic mission and spirit of the national parks system: to be free from commercial exploitation," says former ranger at Yellowstone, Mike Bader. Yellowstone, America's first national park, is home to *Thermus aquaticus*, the microbe that made the DNA-copying polymerase chain reaction possible, and Hoffmann-La Roche billions of dollars. The park service will collect public comments until 15 December and decide by next summer whether to adopt the idea.



Institute allocates \$3bn of stem-cell research spend

When California's \$3-billion largesse for stem-cell research starts flowing in earnest — much of the money is tied up in legal battles — the California Institute for Regenerative Medicine (CIRM) knows exactly how it wants to spend it.

On 11 October the CIRM board discussed draft plans to spend \$1.6 billion on scientific grants, \$487 million on tools, infrastructure and facilities, \$295 million on training and \$220 million on other initiatives.

The plans also lay out specific goals in basic and clinical research. By 2017, the institute hopes to show at least one clinical proof of principle that transplanted cells derived from stem cells help alleviate disease in people. It also hopes to have pre-clinical proof of principle in many animal disease models by then.

Other goals include establishing a new stem-cell bank and gaining a better understanding of what makes stem cells turn into different kinds of adult cells.

the planet was heating up overall and whether climate change could be attributed to the effects of humans.

Protest letters from climate scientists — including Jean Jouzel, director of the Pierre-Simon Laplace Institute, near Versailles — subsequently hailed down on *L'Express*, the French Academy of Sciences and the science ministry. They slammed the assertions as irresponsible given Allègre's public prominence and pointed out errors, such as citing a *Science* paper on Antarctic

ice volumes when it was about snowfall levels (A. J. Monaghan *et al. Science* 313, 827–831; 2006).

On 5 October Allègre stood by his assertions, stating that although he didn't deny climate change, he "considers that global warming was not an essential phenomenon". He dismissed "the fanatics of the greenhouse effect" and said he belonged to the minority who question the link between increasing carbon dioxide levels and climate.

Correction

The News Feature "Shaken, not stirred" (*Nature* **443**, 390–391; 2006) referred to Istanbul as Turkey's capital. The capital is Ankara.

BUSINESS

Hell on no wheels

The crash of a demonstration train in Germany casts a shadow on magnetic levitation technology. **Ned Stafford** reports.

When an advanced prototype train crashed into a maintenance wagon in Germany late last month, 23 passengers were killed. But the accident may have also dealt a mortal blow to the long-touted idea of fast passenger trains that float on magnets, transport specialists say.

Although human error was cited as the probable cause of the accident, the crash has raised safety issues that will now make it much more difficult to gain final approval for a proposed airport shuttle in Munich — a contract vital to the future of Germany's magnetic levitation or 'maglev' train.

Transrapid International, a Berlin-based joint venture between Siemens and ThyssenKrupp, and the Central Japan Railway of Nagoya (JR Central) run the world's two principal projects aimed at developing maglev trains, which glide along almost without friction at speeds of 500 kilometres per hour or more. That compares with a top test speed of 515 km per hour for the world's fastest conventional train, France's TGV — although the TGV travels at only 320 km per hour on service routes.

Between the two of them, Transrapid and JR Central have yielded just one commercial maglev sale so far: Transrapid's contract to build a 30-kilometre airport shuttle in Shanghai, China, which was completed in 2002 at a cost of US\$1.2 billion. Executives from Siemens and ThyssenKrupp say that additional foreign sales will be next to impossible unless a showcase maglev line is built in Germany.

Immediately after the 22 September accident, near Lathen in northern Germany, Munich mayor Christian Ude spoke out against a proposed €1.6-billion (US\$2 billion), 37-km maglev shuttle between the city's train station and its airport, saying he preferred a conventional high-speed train costing just €500 million. He alleged that safety concerns about Transrapid's Munich concept had already been voiced in planning meetings — particularly concerning a proposed 3-km tunnel. Solving the



Broken dreams: the reputation of 'maglev' trains has been damaged by a fatal accident in Germany.

safety problems, he said, could drive costs up to €2.5 billion or more.

Supporters of the project refute Ude's safety claims, and transport specialists around the world are split on whether the crash suggests that the Transrapid is unsafe.

JR Central barely skipped a beat after the accident, announcing just three days later that it would invest ¥355 billion (US\$3.1 billion) more in the technology over the next ten years. It plans to build a larger demonstrator of its

existing design, and to develop a train using superconducting magnets that operate at a higher temperature than before, to reduce the need for magnet cooling. Last November, the company tested its first train using its latest magnets, and achieved a top speed of 553 km per hour.

JR Central spokesman Taro Yoshikawa says an accident like the one in Germany was unlikely with Japan's maglev train because of differences in design. The Japanese train runs

inside a U-shaped double track, whereas the German one sits on a monorail. But Yoshikawa declined to comment on whether the Japanese system is safer overall than Transrapid's.

"There was human error, but one should not have relied on human control. There are definitely safety problems," says Helmut Holzapfel, a civil engineer specializing in transportation at the University of Kassel. One of the major safety issues is the relatively light weight of maglev trains. "Any obstacle on a maglev track presents a problem," he says. "The heavier the obstacle, the bigger the problem — it's just simple physics."

One potential safety measure would be automatic shutdown whenever an object touches the track. But this could be triggered by non-dangerous objects such as birds. Other potential measures include radar devices installed in the train or video cameras mounted along the tracks or in the train.

Transrapid's train has had an uneven ride since research began on it more than 30 years ago. Germany has backed the project heavily in the hope of export sales — but selling it at home has been difficult enough. A proposed 300-km link between Berlin and Hamburg was abandoned in 2000 after eight years of consideration, as was a Dortmund–Düsseldorf link three years later. Transrapid finally succeeded with China — but only after transferring some of its technology to the Chinese and receiving a large but undisclosed support package from the German federal government.

Financial details of the Chinese deal are hard to come by. But ThyssenKrupp's last annual report gives Transrapid sales for 2004 as just €21 million, out of a group total of €5.7 billion. Transrapid accounted for 200 employees out of 27,500.

And just days before the fatal Transrapid accident, ThyssenKrupp chief executive Olaf Berlien indicated that Transrapid would pursue its technology in partnership with the Chinese if the state and federal governments fail to fully approve the Munich contract within 18 months.

John Harding, a physicist and former chief maglev scientist for the US Department of Transportation's Federal Railroad Administration, believes that airport shuttles are not the best application of maglev technology in any case, as short maglev lines don't save much time compared with normal trains. The Munich plan is "very hard to justify", he says.

But he is quick to add that the Lathen crash has not changed his view that maglev technology can be safe. "It was not a technology issue," he says. "They just screwed up."

Additional reporting by Ichiko Fuyono in Tokyo.

"There was human error, but they should not have relied on human control. There are safety problems."

— Helmut Holzapfel

S. FRANKLIN/GETTY IMAGES

THE REAL SEA CHANGE

What can pirates' journals and centuries-old cookbooks teach modern-day ecologists? **Mark Schrope** meets the researchers who trawl history books for deeper insights into marine ecosystems.

In the late 1600s, Alexandre Olivier Exquemelin was busy living up to his extravagant name. In his book *The Buccaneers of America*, the French-born pirate describes a host of battles and "barbarous inhumanities" through the Caribbean and along the South American coasts. He writes about some pirates' habit of taking Cuban sea-turtle fishermen as slaves, and he personally owns up to stealing victuals left ceremonially by grieving widows at their husbands' graves.

But between all the sacking and pillaging, careful readers can also find glimpses of natural history. A description of a green sea turtle reads: "Their eggs... are found in such prodigious quantities along the sandy shores of those countries that, were they not frequently destroyed by birds, the sea would infinitely abound with tortoises... Certain it is that many times the ships, having lost their latitude through the darkness of the weather, have steered their course only by the noise of the tortoises swimming that way, and have arrived at these isles."

Today, a growing number of marine scientists argue that researchers focus too narrowly on recent decades and don't make proper use of historical records to put their work in context. Clearly, the most recent data are also the most thorough. But some argue that findings from such studies are improperly informed at best, and at worst support fundamentally flawed management schemes. What if, for instance, an ecosystem had been severely degraded a century earlier, but the general understanding of the ecosystem — including restoration targets for fish and other marine creatures — was based on observations made over only a few decades?

Time team

One way to resolve this is to look at the past, and detectives in the emerging field of historical marine ecology are doing just that, scouring archives, museums and archaeological and palaeontological records. They root through relatively straightforward data, such as how many fish were caught in a particular harbour. But they also have to get creative, delving into obscure sources such as pirate logs, medieval cookbooks and restaurant menus that track the availability of seafood.

Most historical marine-ecology projects fall under the Census of Marine Life, a ten-year global initiative with the ambitious goal of cataloguing everything that lives in the

ocean. An arm of the programme — History of Marine Animal Populations (HMAP) — aims to determine what used to live there. By 2010, researchers will have examined more than a dozen key sites around the globe; historical studies are already under way in southwest Africa, Australia, Europe and the United States, including the latest, a study of the Florida Keys. Further work is planned in southeast Asia and New Zealand.

HMAP was designed to correct the "historical myopia" of fisheries scientists, says Tim Smith, one of the programme's leaders and a fisheries biologist retired from the National Marine Fisheries Service in Woods Hole, Massachusetts. In 1995, a seminal paper appeared describing the basic problem as "shifting baseline syndrome"¹. In it, Daniel Pauly of the University of British Columbia argued that researchers base their understanding of healthy

fish populations on what fish stocks are like during their lifetimes, with no framework for incorporating how much more plentiful fish might have been in generations past.

Although historical documents can shed light on this, there are limits, says David Starkey, another programme leader and historian at the University of Hull, UK. "You can never revisit the past and replicate what went on — you can only get glimpses of the past," he says. "That's an inevitable challenge." But not facing that challenge is a mistake, he believes.

The one-year Florida Keys project began in June and is designed to shift baseline observations of the ecosystem as close as possible to their positions before Europeans showed up. Graduate student Loren McClenahan, working with marine ecologist Jeremy Jackson at the Scripps Institu-

tion of Oceanography in San Diego, California, and others, is combing Spanish and British archives for records dating back to the 1600s, and US libraries for accounts from the 1800s onwards. She hopes to learn how far coral reefs used to extend around the Keys, as well as the population history for 14 important species including mangrove trees, grouper fish and the now-extinct Caribbean monk seals.

Rogues' gallery

Some sources are richly detailed. In 1803, US commissioner Andrew Elicott wrote in his journal: "Along the Florida Reef, and among the Keys, a great abundance and variety of fish may be taken: such as hog-fish, grunts, yellow tails, black, red, and gray snappers, mullets, bonefish, amber fish, margate-fish, baracoota, cavallos, pompui, groopers, king-fish, siber-fish, porgys, turbot, stingr, black drum, Jew fish, with a prodigious variety of others, which in our situation we found excellent."

In an earlier related study², McClenahan and Jackson looked at historical populations of sea turtles throughout the Caribbean. They found the late-seventeenth-century writings of a buccaneer named William Dampier particularly useful. For instance, he described the location of turtle nesting areas and aggregations in enough detail for modern scientists to map them accurately.

Pirates may seem notably suspect as sources of information. Many pirate accounts exaggerate their exploits, says Starkey, because interesting tales from their world travels could make them money and also deflect attention from their illegal activities. But Dampier was one of the most scientifically significant and reliable, says Starkey, who has studied privateers, pirates and buccaneers. And later voyagers often confirmed Dampier's observations.

Even Charles Darwin recognized the value of Dampier's work. He carried some of the

"You can never revisit the past and replicate what went on, you can only get a glimpse. That's an inevitable challenge."
— David Starkey



Tales of the sea: pirate journals dating back to the late 1600s can offer useful information on marine ecosystems.

BRIDGEMAN ART LIBRARY



COURTESY OF MONROE COUNTY PUBLIC LIBRARY, KEY WEST

A question of scale: archive material, such as this photo taken in Florida in the 1940s, can give fresh perspective on the abundance of certain species.

buccaneer's century-old writings on his voyage on *The Beagle*, and referenced them in his account of the trip. He described Dampier as "so accurate a person", noting that a volcanic peak in the Galapagos Islands that he found overgrown must previously have been barren, simply because Dampier described it as such. Even some of Darwin's early work on evolution and natural selection was inspired in part by Dampier's thoughts on "bastard" species, such as Galapagos sea turtles, that he thought must be a mix of species from different geographical areas.

Net gains

In Florida, some of the best descriptions of the historic Keys come not from pirates, but from surveyors, captains and doctors. Not surprisingly, the sources paint a collective view of the Keys quite different from today. In 1796, Scottish surveyor George Gauld wrote that "among the roots of the mangroves and about every old log or piece of rotten work, there are such quantities of the largest crayfish [spiny lobsters] that a boat may be loaded with them in a few hours". Today, divers spend hours filling just one bag. Other accounts spoke of plentiful fish and coral reefs where now only sand or rubble are found.

Tracking down such stories takes a combination of creative thinking and patience. "It's a little bit of a treasure hunt," says McClenachan.

Her greatest find to date is a 1775 map by Gauld that includes details about coral reefs to avoid and views of the approaches to specific islands, giving a snapshot of the area covered by mangrove. "It's just a complete gem," she says.

Based in part on similar analyses of historical sources, Jackson and his colleagues have concluded that coral reefs are in severe decline worldwide, largely as a result of human activities such as overfishing³. Such far-reaching extrapolations have fuelled an ongoing debate over reef health. Richard Grigg, a coral-reef biologist at the University of Hawaii at Manoa, says that although he favours the historical-ecology approach in general, he thinks Jackson's conclusions are too broad. "You have to be extremely careful with that kind of information," he says. "I think it's up to us as scientists to be extremely critical and not fall into sweeping generalizations about how the sky is falling."

For his part, Jackson argues that historical analyses can provide powerful tools for quantitative assessments. Anecdotal evidence, too, can be scientifically valuable, he says. If, for instance, records suggest that sea turtles in a region were once plentiful but are now scarce, that provides important ecological information

about the past — even without details of exactly how many there were.

Like the Keys project, an HMAP study led by Heike Lotze, a marine ecologist at Dalhousie University in Halifax, Nova Scotia, has taken a broad view of a complex ecosystem — in this case the Wadden Sea in northwestern Europe⁴. "If you go there today, most of the area is protected as a national park, so I

think a lot of people envision it as a natural ecosystem," she says. But "500 or 1,000 years ago it was a completely different system. If you have this historical knowledge, you see how much humans have transformed it."

During medieval times, fish populations in rivers and lakes connected to the

Wadden Sea began to drop as people overfished it and built dams, which blocked the spawning migrations. Later dyke building also cut off estuarine nursery areas. Archaeologists have documented how fish populations plummeted — for instance, digs of waste piles from old fish markets show that sturgeon scales became increasingly less common. But cookbooks also tell part of the tale. Sturgeon was once popular even for common folk, but by the fourteenth century they were so rare that they were only served at the king's table. Later

"It's up to us as scientists to be critical and not fall into sweeping generalizations about how the sky is falling."

— Richard Grigg



Archaeological digs of waste piles near nineteenth-century sturgeon fisheries, such as Hamburg harbour, shed light on the fish's dwindling numbers.

French cookbooks held clues to the species' drop in numbers, such as instructions on how to prepare veal as a substitute in recipes calling for unavailable sturgeon.

As the more easily accessible freshwater fisheries collapsed, medieval fishers turned to coastal waters and then moved farther afield as fish populations there dwindled. The Wadden Sea isn't the only place where overfishing hit hard, of course. A team led by Andrew Rosenberg, a marine ecologist at the University of New Hampshire in Durham, has done extensive research to determine historical cod populations off the coasts of the United States and Canada.

Record catch

In one study, for instance, the group calculated the cod population on Canada's Scotian Shelf in 1852, based on detailed logs for schooners operating out of Beverly, Massachusetts⁵. The researchers think the information is reliable because the reporting system offered little incentive for fishers to misreport, and because there is some overlap among the logs. Overall, the study suggests that the 1852 cod population was two orders of magnitude larger than today.

And more than four-and-a-half centuries before the 1990s collapse of the Atlantic cod fisheries, there was an astonishingly rich population. In 1623, fisherman Emmanuel Altham wrote in a letter to his brother: "In one hour we got 100 great [very large] cod... and if we would have but stayed after the fog broke up, we might quickly have loaded our ship... I think we got 1000 in all... When we had nothing to do my

people took great delight in catching them, although we threw them away again."

Rosenberg says it is too soon to say whether fisheries managers will apply such historical records to their efforts. But Thomas Hill, acting chairman of the New England Fishery Management Council and a long-time fisherman based in Gloucester, Massachusetts, believes fisheries management in his region has indeed been hampered by a dependence on modern data. "I think any historical perspective is certainly important in developing long-term public policy," he says. "In the absence of that, you're shooting in the dark." Hill adds that a better understanding of the historical populations could help fishermen appreciate their collective impact.

One HMAP study that could have controversial management implications involves estimates of world whale populations. Smith has looked through sources such as customs-house records and period trade journals to estimate global catch rates for whale species including

the humpback and the North Atlantic right⁶. The problem is that newer genetic analyses estimate total population sizes in the 1800s to be at least an order of magnitude larger than those based on historic records⁷. In this case, pirate activities may have complicated the historical analyses, as whale catches for ships they captured were not recorded. Known sources of underreporting, though, would likely change estimates only by a few per cent, says Smith, and work continues to reconcile the discrepancy.

Despite such controversies, the HMAP researchers believe that their work is gaining momentum and acceptance. "The hybridization of history and ecology has really put a set of new questions on the table that, together, the scholarly community has been able to answer," says Jesse Ausubel, an ecologist at Rockefeller University in New York and programme director for the Census of Marine Life at the Sloan Foundation. And overall, says Jackson, the work has potential to make a real impact. "It's a whole new way of viewing the world," he says. "If in some small way this can shape ecological research for the future and help us to conserve biodiversity, then that will be wonderful."

Indeed, if Darwin is any indicator, even a pirate could help shape that future.

Mark Schroppe is a freelance writer in Florida.



Artists' impressions from Scandinavia (c. 1555) indicate an abundance of river fish populations.

1. Pauly, D. *Trends Ecol. Evol.* **10**, 430 (1995).
2. McClenachan, L., Jackson, J. & Newman, M. *Frontiers Ecol. Environ.* **4**, 290–296 (2006).
3. Pandolfi, J. M. et al. *Science* **307**, 1725–1726 (2005).
4. Lotze, H. et al. *Helgol. Mar. Res.* **59**, 84–95 (2005).
5. Rosenberg, A. et al. *Frontiers Ecol. Environ.* **3**, 84–90 (2005).
6. Smith, T. & Reeves, R. *Caribbean J. Sci.* **39**, 286–297 (2003).
7. Roman, J. & Palumbi, S. R. *Science* **301**, 508–510 (2003).

A question of trust

AIDS treatment in South Africa is often a tug-of-war between clinicians and traditional healers. **Natasha Bolognesi** meets a woman who is uniquely qualified to heal the rift.

For most people in the developed world, the words 'traditional healer' conjure up the image of a figure cloaked in beads, animal pelts and an air of impenetrable mystery. Someone, in short, whom Westerners find difficult to understand or trust, and who has rejected biomedical science in favour of mysticism and magic.

This kind of distrust is problematic anywhere, but especially so in African countries struggling with the HIV epidemic. South Africa, the worst affected, is home to an estimated 5.5 million HIV-positive people, many of whom visit traditional healers, or *sangomas* as they are known locally. Many *sangomas* have earned themselves a bad reputation among doctors for not referring their HIV patients to clinics for testing and treatment. For their part, many traditional healers fear that Western medicine will harm their patients.

Amid this welter of misunderstanding, there is one *sangoma* who finds herself in the unique position of being able to understand and relate to both sides of the divide. She is finding ways to reconcile the two and to build bridges of trust between doctors and *sangomas*. And now, her efforts are enriching and informing a pilot project to improve the health and quality of life of HIV patients in South Africa.

British born and bred, Jo Wreford is a doctor of social anthropology and a research fellow for the AIDS and Society Research Unit at the University of Cape Town. She is also one of a small number of white people in South Africa who have qualified as *sangomas*, and is known to her *sangoma* colleagues and clients as Thobeka, which means 'she who can be trusted because she is grounded' in the Xhosa language. This training, which involved undertaking a

spiritual journey with a *sangoma* mentor, makes it easier for other healers to trust her as one of their own. Her training in anthropology and Western cultural background reassures doctors that she understands and appreciates scientific method.

Different vision

Wreford wears beads, throws the bones, burns the herb imphepho to invoke the guidance of the ancestors, and experiences visions. She says that unlike many of her colleagues, she is comfortable with reconciling her traditional beliefs with science, arguing that biomedicine can treat the body, while traditional healing can help treat the soul. "Neither cures," she says. "We must therefore use the best of what both systems have to offer to alleviate the AIDS burden, which is both physical and psychological."

Wreford wants to see doctors and traditional healers working harmoniously side by side. "Both play vital roles in healing the majority of AIDS-afflicted South Africans," she says. "ARVs (antiretrovirals) are the only medical intervention available to alleviate the physical effects of AIDS. The traditional healer, in addition to using herbs, also works on the spiritual level, which is an essential part of the African healing process that Western medicine does not address." But this will be a huge challenge. Del Kahn of the department of surgery at the University of Cape Town explains the obstacles: "Most doctors still regard *sangomas* with suspicion because they don't have training in treating serious organic disease, and the general

feeling is that traditional healers do more harm than good in patients with organic illness."

Sangomas are just as wary of Western physicians. Phillip Kubukeli, founder and president of the Western Cape Traditional Healers and Herbalists Association based in Cape Town, says that during the apartheid regime in South Africa many *sangomas* believed that doctors administered poison instead of medicine to black patients in the hope of killing them off. "This fear persists today," says Kubukeli, "although it is not as prevalent as it was."

Kubukeli adds that traditional healers strongly believe that physicians are out to get their hands on herbal remedies to sell to drug companies. "Many Western medicines are derived from indigenous plants," says Kubukeli. "This makes traditional healers very suspicious. For example, when I am trying to bridge the gap between *sangomas* and doctors, the *sangomas* will often accuse me of trying to sell their remedies to the physicians."

Driving a deeper wedge between this mutual medical divide is the South African government's perception that natural remedies can treat AIDS — a view it vigorously promotes. At the International AIDS Conference in 2000, South African President Thabo Mbeki caused an international uproar when he questioned the link between HIV and AIDS. And at this year's AIDS Conference in Toronto — where South Africa's stand, covered in beetroot, garlic and lemons, seemed seriously out of place — the South African health minister Manto Tshabalala-Msimang further discredited her government's AIDS policy by saying people must have a choice between ARVs and traditional remedies (see *Nature* 443, 134–135; 2006).

"As a result of this persistent denial, even at the highest government levels," says Monika Esser, a paediatrician at Tygerberg Academic Hospital in Cape Town, "Western medicine of predominantly white origin continues to be met with an element of suspicion by black patients and traditional healers."

It is here, into the turmoil of the AIDS healing conflict in South Africa, that Wreford hopes to throw a lifeline, in the form of a willingness

"Most doctors still feel that traditional healers do more harm than good in patients with organic illness."

— Del Kahn

S. HIPPLER



Body and soul: to Jo Wreford (centre and right), clinics and healers are both needed to fight AIDS in Africa.

N. BOLOGNESI

to share her acceptance of both scientific and traditional beliefs with doctors and *sangomas*. Her hope is that by trusting Thobeka, the one who can be trusted, they can learn to trust each other.

Wreford qualified and practised as an architect in London before deciding to explore the spiritual aspect of her life more, and use this to help people through spiritual healing in Africa. Before leaving London, she embarked on a course of jungian psychotherapy, which emphasizes exploring the unconscious through the use of visualization and dreams. Wreford says it helped to prepare her for the intense spiritual demands made on her during her *sangoma* training.

What Western medicine needs to understand, says Wreford, is that many Africans believe that their ancestors live in a separate realm and carry with them answers to the deep questions about the cause of illness. "This knowledge is accessed by the *sangoma* through ritual, visions, dreams and herbs, and communicated to the patients, who then feel they have redressed the situation, which prompts more complete healing," she explains. "Like psychotherapy, it can also help people cope with stigma and emotional strain in the face of a disease such as AIDS."

Spreading HOPE

Wreford plays an equally vital role in explaining to her *sangoma* colleagues the importance of biomedicine in fighting AIDS. "Here my role is specific," she says. "*Sangomas* believe that the disappearance of symptoms through the administration of a herb means that they have cured the patient of AIDS. I have to explain to them that this is not so and that the application of ARVs, which do not cure either, at least enables the patient to live life normally."

But alone, her efforts are a drop in the ocean. According to the Western Cape Traditional Healers and Herbalists Association, there are more than 200,000 *sangomas* in South Africa — vast numbers spread over a vast country.

So Wreford was invited to join forces with HIV Outreach Programme and Education (HOPE), a non-profit, non-governmental organization based in Cape Town. HOPE is currently running a pilot study in five townships outside Cape Town to build mutual trust and acceptance between *sangomas* and doctors. It aims to encourage them to collaborate on HIV/AIDS intervention, to avoid disruption



Wreford uses herbs, visions and ritual to access spiritual knowledge for clients.

of ARV treatment through mistaken herbal administrations and to persuade more male clients to volunteer for HIV testing.

The project operates on a cross-referral system between nine traditional healers and primary-health-care clinics in the townships. The healers, having been taught to recognize the symptoms of HIV infection and the importance of ARVs, refer these patients to a clinic for testing, counselling and treatment. The patients then come back for traditional spiritual counselling from the healer.

HOPE's recruitment and training of *sangomas* for the pilot study in October 2005 sparked misgivings at first. According to HOPE training officer Pauline Jooste, one healer said: "We were very scared — the facilitator was a white person." Now the response seems more positive. Nomsisi Stefans, a practising *sangoma* from Mfuleni township outside Cape Town, says: "I am happy with HOPE and, if I think they should, my clients are happy to go to the clinic to be tested, especially if I go with them."

The project is closely monitored and aided by Wreford. She participates in a monthly support and supervision day for all participants, regularly visits the clinics and *sangomas* to

ensure commitment and quality, and lectures medical students and doctors on the need to recognize and respect the value of spiritual healing.

Much help is necessary if the project is to work. Nocawe Frans, a HOPE member and social worker at Tygerberg hospital, points out that many traditional healers are reluctant to spread HOPE's message because they resent Western interference in African traditional medicine. "It is also difficult for *sangomas* to refer patients to a hospital because in African culture the hospital carries a strong association with death — parents often take their children out of hospital and to a *sangoma* instead," she says.

Despite these challenges, the five-month-old project is beginning to show a slight increase in patient referrals to clinics from traditional healers, although exact numbers are hard to come by, says Esser, who is also a HOPE management member. Patients tend to go to clinics outside their community owing to the persistence of an enormous AIDS stigma in South Africa.

Also encouraging are signals from both camps that Wreford's and HOPE's efforts are working. "I think we underestimate the spiritual needs of patients — the 'healing' rather than the management of treatment and cure that we are familiar with in Western medicine," says Helena Rabie, a specialist in HIV/AIDS in children at

Tygerberg hospital. "If we can succeed in tapping into the traditional healers' influence as a resource to fight the spread of HIV, it would be wonderful."

Confidence is rising among *sangomas* too, according to Kubukeli. "Thobeka is truly great," he says. "She has helped my colleagues understand Western medicine. We want to see the HOPE project grow."

HOPE chairman, Reverend Stefan Hippler, is optimistic that, if the project succeeds, it could inform HIV policy in other African nations as well. "Traditional health practitioners are important role models and leaders in their communities," he says. "They are indispensable for all national efforts in the fight against HIV and AIDS."

Natasha Bolognesi is a freelance health writer based in Cape Town.

"It is difficult for healers to refer patients to hospitals, which in Africa carry a strong association with death."

— Nocawe Frans



WARNER BROS./THE KOBAL COLLECTION

Comfortably numb

It started life as an anaesthetic, then became a psychedelic club drug. Now researchers think ketamine could hold the key to understanding and treating depression, says **Erika Check**.

B held my hand and we started our roller coaster out... I can't feel my body anymore except this overriding general fuzziness. The lines on the ceiling become a tunnel and I am flying down it faster than sound... Then the room does a somersault and I with it... I'm scared... this is a thrill ride and the car is the dimensions of existence.

Such are the wonders of the drug ketamine, according to U.S. 9, who described this experience with the drug on an online forum called the Erowid Experience Vaults¹. Many others around the world use ketamine illegally, going “down the K-hole” to abandon reality and alter their consciousness. Now neuroscientists are getting in on the act. They are finding that although ketamine makes some lose their minds, it might help others to find their sanity.

Ketamine was invented in the 1960s by chemists at the drug company Parke-Davis in Detroit. The drug was a powerful anaesthetic, but also caused ‘dissociative’ effects, such as the feeling of leaving one’s body and entering other planes of existence. Although popular with recreational drug users, ketamine is only used medically as an anaesthetic in animals and children, who are less prone to its psychedelic side effects.

But back in the 1960s, a handful of scientists used these side effects to explore human consciousness (see ‘Exploring the dark’, overleaf). Today, scientists are using ketamine as a tool to study mental illness, and perhaps even treat it. Clinical trials suggest that the drug might help patients with severe depression. And these experiments are changing the way scientists think about the nature of the disease and how it might be treated.

The World Health Organization estimates that depression is the world’s leading cause of disability. But doctors and scientists are still mystified by what the disease actually is. The label ‘depression’ is a catch-all term for a condition with a huge array of possible symptoms and causes. Some depressed people feel sad, guilty or tired; others cycle into hyper-excited mania, or feel worried and anxious. Genes, stress and negative thought patterns have all been linked to depression. But it is not clear what chain of events causes a particular set of symptoms in an individual.

Scientists also aren’t quite sure why modern antidepressant drugs succeed or fail to cure depression in different patients. The drugs act on neurotransmitters, the chemicals that brain

cells use to communicate. Most of today’s drugs target a particular class of neurotransmitter called the biogenic amines. These include serotonin, which is targeted by selective serotonin reuptake inhibitors, or SSRIs. Fluoxetine and sertraline, sold in the United States as Prozac and Zoloft, are members of this enormously popular category of drugs.

But the drugs that act on biogenic amines take weeks to work, and fail to help at least 40% of depressed patients². So neuroscientists suspect that these drugs don’t hit depression at its source and are searching for other approaches.

The search took a dramatic turn in August this year, when a team led by Husseini Manji, at the National Institute of Mental Health in Bethesda, Maryland, published a study³ looking at the effects of ketamine on 18 severely depressed patients, all of whom had failed to respond to standard treatments.

In the double-blind trial, doctors gave the patients intravenous ketamine or placebo saline drips, and then scored the responses. After taking ketamine, 12 of the patients improved by at least 50% on a depression rating scale. Patients felt better as little as two hours after treatment.

It’s surprising. If anything, I would have expected these patients to crash and burn.”

— Nuri Farber

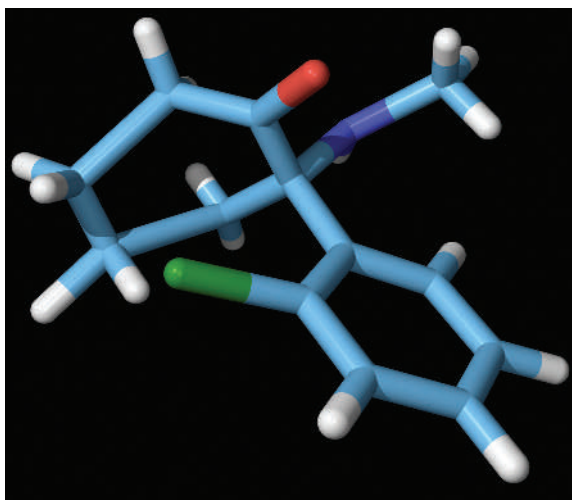
And one-third of the patients still felt better a week later.

Although unexpected, Manji's results are not the first to hint that ketamine has unexpected benefits. In 2000, John Krystal at Yale University and his colleagues published the results of a similar trial on eight severely depressed patients who didn't respond to standard medications⁴. This too suggested that ketamine might work as an antidepressant. Four of the patients felt much better after the ketamine treatments — their depression scores dropped by more than half, by one rating scale. And it worked fast. The patients felt better just three days after their treatment.

Special forces

At the time, these results astonished other psychiatrists. "Ten years ago, I would have said there's no way in hell this drug would work as an antidepressant," says Nuri Farber, a psychiatrist at Washington University in St Louis. Many remained sceptical, and were wary about the whole idea of treating mentally unstable people with a psychoactive drug. Psychiatrists have on occasion used ketamine to deliberately destabilize normal subjects: by giving it to stable people, psychiatrists can induce a schizophrenia-like state and study the brain chemistry that may underlie the disease⁵. So psychiatrists didn't exactly rush to try ketamine in their own clinics. "It was such a small study, and nobody else was following this path. It was sort of a novelty," Krystal says. "People didn't take it as seriously as they might have."

The findings languished in the literature until one of Krystal's colleagues, Dennis Charney, left Connecticut to work at the National Institute of Mental Health. There, he began working with Manji, who had been studying chemical relatives of ketamine in mice. Manji, Charney and a psychiatrist named Carlos



In the frame: ketamine may ease the effects of stress on brain cells.

Zarate decided to find out whether the ketamine study was more than a fluke.

When they took ketamine, Zarate's patients did experience trippy side effects such as dizziness and euphoria. But most of these effects wore off after 80 minutes³ — before the drug's antidepressant effects kicked in. That is surprising: usually, drugs that trigger highs, such as cocaine or ecstasy, are followed by a depressive low. "It's not what I would have expected," says Farber. "If anything, I would have expected these patients to crash and burn."

One possible explanation for this is that ketamine uses different pathways to trigger its psychedelic and antidepressant effects. Another possibility is that patients have to go out of their minds before they can get back to normal. "What ketamine does is briefly make people crazy," says Eric Nestler, a neuroscientist and psychiatrist at the University

of Texas Southwestern in Dallas. "The question becomes: is that disruption in cognitive function what's creating this improvement in mood?"

Right now, it is impossible to answer that question. But Zarate, Krystal and other researchers who have studied ketamine's link to depression have one idea about what's going on. Their hypothesis has to do with the way ketamine works in the brain.

Ketamine hinders the activity of a complex molecular machine called the N-methyl-D-aspartate receptor, the NMDA receptor for short. It is one of the receptors for a neurotransmitter called glutamate. That activity could relate to one theory of depression: that the disease occurs when brain cells are just too stressed to thrive. This link is based on two facts and one highly controversial idea.

First, scientists know that the NMDA receptor can control brain-cell growth and survival. Second, they know that excessive levels of glutamate kill brain cells in some conditions, such as stroke and Alzheimer's disease⁶. Here's the controversial part: there is some evidence supporting the idea that depression is caused by brain-cell death.

So, by jamming the NMDA receptor, could ketamine be correcting a toxic glut of glutamate that is harming brain cells and causing depression?

It is already known that stress floods the brain with glutamate, says Zarate. "It might be that these neurons are struggling to regulate glutamate, and if you stress them over and over, they become injured."

This hypothesis is still very new⁷. There is some evidence linking glutamate, and the NMDA receptor, to depression. Twenty-five years ago, for instance, scientists at the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland, showed that chemicals that target the NMDA receptor have antidepressant effects in animals⁸. Scientists have since found that deceased depressed human patients, such as suicide victims, have abnormal numbers of NMDA receptors. And brain-imaging studies have found that depressed people have much higher levels of glutamate in one region of their brains than healthy people⁹.

But Nestler and other psychiatrists caution that it is impossible to know yet whether ketamine affects brain-cell survival. Glutamate is the most common neurotransmitter in the brain, used by perhaps half of all brain cells. And the NMDA receptor is involved in a huge array of different processes, such as learning and memory, as well as cell growth and survival. So it is difficult to pin down the precise reason why tweaking glutamate through the NMDA receptor would influence human

"What ketamine does is briefly make people crazy. Is that what's creating this improvement in mood?"

— Eric Nestler



John Krystal published one of the first studies to use ketamine as an antidepressant.

T. DAGRADI

happiness. And although it is true that the NMDA receptor is involved in cell survival, this takes a long time, whereas ketamine's antidepressant effects seem to kick in within hours.

Until the mechanism can be clarified, psychiatrists say, Zarate's and Krystal's studies are important for one major reason: they provide evidence that biogenic amines such as serotonin don't tell the whole story about depression. "We've had a preoccupation over so many years with biogenic amines," says John Olney, a psychiatrist at Washington University who has been studying neurotransmitters for 35 years. "The idea that glutamate might be involved in depression has evolved very slowly. We're still trying to understand it."

What's the use?

Nestler agrees. "All our available drugs act on the serotonin and noradrenaline systems, and this drug clearly does not," he says. "It's very important as a proof of principle that a drug acting on a different neurotransmitter system can have a mood-elevating effect."

Even if ketamine works, it's not an ideal drug. Clinical trials have studied the drug in only 26 patients, and no one has investigated how long its beneficial effects might last; there are also those psychedelic side effects. So neuroscientists are continuing to look into whether other drugs that hit the glutamate system could help mentally ill patients. A trial of one NMDA blocker — memantine, used to treat Alzheimer's patients — found the drug didn't cure depression¹⁰. But the study's authors think that is because memantine does not bind as tightly to the NMDA receptor as ketamine does.

Another candidate, riluzole — already used in patients with Lou Gehrig's disease — seems more promising. It works by preventing brain cells from making excess glutamate and dumping it into surrounding brain tissue. Patients taking the drug have reported some improvement, but riluzole has yet to undergo large clinical trials for depression. Other drugs are entering tests.

Whether or not these trials prove successful, at least the fresh take on depression is finally helping neuroscientists climb out of the serotonin rut. And maybe an ageing drug with a tarnished reputation will help them find their way.

Erika Check is a reporter for Nature in San Francisco.

1. www.erowid.org/experiences/exp.php?ID=1959
2. Berman, R. M., Narasimhan, M. & Charney, D. S. *Depress. Anxiety* **5**, 154–164 (1997).
3. Zarate, C. et al. *Arch. Gen. Psychiatry* **63**, 856–864 (2006).
4. Berman, R. M. et al. *Biol. Psychiatry* **47**, 351–354 (2000).
5. Abi-Saab, W. M., D'Souza, D. C., Moghaddam, B. & Krystal, J. H. *Pharmacopsychiatry* **31** (Suppl. 2), 104–109 (1998).
6. Sonkusare, S. K., Kaul, C. L. & Ramarao, P. *Pharmacol. Res.* **51**, 1–17 (2005).
7. Kugaya, A. & Sanacora, G. *CNS Spectrums* **10**, 808–819 (2005).
8. Trullas, R. & Skolnick, P. *Eur. J. Pharmacol.* **185**, 1–10 (1990).
9. Sanacora, G. et al. *Arch. Gen. Psychiatry* **61**, 705–713 (2004).
10. Zarate, C. A. et al. *Am. J. Psychiatry* **163**, 153–155 (2006).
11. Lilly, J. C. *The Scientist: A Metaphysical Autobiography* (Ronin, Berkeley, CA, 1996).



J. BRYSON/TIME LIFE PICTURES/GETTY

Exploring the dark

"You're supposed to be reputable scientists! Not two dorm kids freaking on Mexican mushrooms!"

The 1980 movie *Altered States* tells the story of a scientist called Eddie Jessup who uses psychedelic drugs to explore the depths of human consciousness. A fellow scientist objects to the experiments, chastizing Jessup when he witnesses the outlandish behaviour his drug-taking produces.

The Jessup character was partly inspired by reality: a number of scientists in the 1950s and '60s experimented on themselves, using drugs, including ketamine, to explore the nature of consciousness.

Among them was John Lilly, who worked at the National Institute of Mental Health in the 1950s. While

at the institute, Lilly began working with a sensory-deprivation tank (pictured). After leaving the institute in 1958, Lilly embarked on a full-time career in self-experimentation with his own sensory-deprivation tank and a cornucopia of psychedelic drugs, including LSD (lysergic acid diethylamide) and ketamine. Scenes from *Altered States* are actually based on Lilly's experiments with ketamine, which he described in his autobiography *The Scientist*¹¹.

Lilly began his scientific life with mainstream studies in electrophysiology, but his later career took him down some bizarre paths. He began working extensively with dolphins, even dosing them with LSD, believing that humans could communicate

with them. And he began an extensive programme of experimentation with ketamine. He believed the drug connected him with aliens, and with a God-like entity he called the Earth Coincidence Control Office.

But some of his beliefs weren't quite so distant from the scientific mainstream. In a 1991 interview with the editors of a book called *Mavericks of the Mind*, Lilly said that he agreed with the idea that "the brain is a huge, diverse chemical factory". Lilly told his interlocutors: "We cannot make generalizations about any one of these chemicals yet but, for instance, if you give an overdose of this one people get depressed, if you give an overdose of that one they get euphoria, and so on." **E.C.**

Biochemistry has outgrown its traditional boundaries

SIR — Your Editorial “What’s in a name?” (*Nature* **442**, 486; 2006) makes the case for the importance of chemistry to current ‘molecular sciences’ but contains the unfortunate statement that biochemistry is “associated primarily with the study of enzyme kinetics.” This is a narrow and dated view of the subject.

In fact, the revolution in modern biology has been underpinned by the application of traditional biochemistry to fields as diverse as genetics, cell biology and ecology. To acknowledge the importance of chemistry to biology but to regard biochemistry as little more than the study of enzyme kinetics displays an ignorance both of the courses that make up a modern degree in biochemistry and of the content of biochemistry journals.

What is in a name? Many traditionally trained biochemists would happily describe themselves as molecular or cellular biologists. The boundaries in science have all but disappeared, and a good thing too. However, if we are to debate traditional labels, please don’t take the richly diverse subject of biochemistry and define it by a single discipline.

Chris Kirk

The Biochemical Society, Eagle House, 16 Procter Street, London WC1V 6NX, UK

Videos have starring role to play in protocol sharing

SIR — Who has never struggled with a new protocol? The electronic initiatives to share and discuss protocols, as reported in the News story “Online methods share insider tricks” (*Nature* **441**, 678; 2006) and in Correspondence (“Wiki and other ways to share learning online” *Nature* **442**, 744; 2006), are to be welcomed. Unfortunately, even a detailed written protocol cannot have the same effect as sitting beside someone who is doing it.

Probably the most feasible approach is to publish movies describing the methods, a service already offered by some publications and protocol websites, but which could become routine. Much more information on the essential steps of a new protocol, including audio commentary on the trickiest steps (from the position of the Petri dishes to the speed of dispensing), could be accessible using video format and published online with the paper. Such videos could transform the way in which methods are communicated.

Readers may object that they don’t have time to set up Hollywood studios to publish a paper. But the effort required for filming a relevant procedure is outweighed by the fact

that more precise description of methods greatly aids scientific understanding and facilitates future research. Furthermore, videos are valuable tools for teaching.

Matias Pasquali

Centre of Competence for Innovation in the Agro-environmental Sector, University of Torino, and Department of Plant Pathology, 1991 Upper Buford Circle, University of Minnesota, Saint Paul, Minnesota 55108-6030, USA

Consistency tests establish empirical generalizations

SIR — Your Editorial “Let’s replicate” and News Feature “The trouble with replication” (*Nature* **442**, 330 and 344–347; 2006) address the problem of how to obtain reliable knowledge in science. These articles usefully point to difficulties in conducting successful replications, and note the change in social processes associated with resolving contested experimental results. Taken together, though, they risk misleading the reader about the process of replication in science.

Your Editorial acknowledges the diversity of the practice of replication in science, but uses a rather narrow definition of replication. By writing of “direct”, “full” and “exact” replication, you seem to presuppose an ideal of literal replication — the idea that replication involves attempts to produce a ‘carbon copy’ of the original experiment. Of course, such an objective cannot be attained, if only because of the unavoidable changes in time, location and investigative personnel that replication brings with it.

Close replication is a more realistic aim in science than literal replication. A close replication uses the conditions and procedures of the initial study, perhaps correcting for its perceived inadequacies. It is a checking strategy that attempts to achieve similar results to the first study.

Close replication must, however, be distinguished from constructive replication. The latter process is undertaken to demonstrate the extent to which results hold across different experimental methods, treatments, populations and occasions. It is a triangulation strategy designed to reveal the extent to which the results identified by successful close replication can be generalized. Constructive replication is of vital importance to science because it is the main means by which empirical phenomena are detected. Your Editorial mentions the scientific aim of confirming and extending unpublished experimental conclusions, but in my view it errs in not stating that this is a form of replication.

To be sure, craft or tacit knowledge is often needed to establish the reliability associated with effectively carrying out experimental studies and successfully replicating their

findings. Social studies of science have made clear that there is also an unavoidable social dimension to the successful practice of experimental science. However, it is the successful use of consistency tests in the practice of experimental replication that provides the evidential basis for its claims about empirical phenomena.

Brian D. Haig

Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Chimp comparisons won’t explain human evolution

SIR — You state in your News story on genetic differences between humans and other species (“Mix and match: the hunt for what makes us human” *Nature* **443**, 8; 2006) that research is beginning to pin down genes that “evolved rapidly during the transition from chimps to people”. No such transition occurred, of course, because chimpanzees are not human ancestors; they have been evolving for exactly the same amount of time.

A mutation is, in principle, just as likely to have occurred in the chimpanzee lineage as in the human lineage, during the time since their common ancestor lived. There is no way to discover from any comparison of the two species which is the case.

To determine the evolutionary history of a trait and its genetic basis, it is necessary to make phylogenetic reconstructions based on comparisons among many species. Chimp–human differences make good headlines, but tell us little about how human uniqueness evolved.

Robert Barton

Evolutionary Anthropology Research Group, Durham University, 43 Old Elvet, Durham DH1 3HN, UK

Science prediction rate would be good in gambling

SIR — I was surprised to read in Don Ihde’s review of David Nye’s *Technology Matters: Questions to Live With* (*Nature* **442**, 984; 2006) that “of the 1,500 published predictions from scientists, inventors and sociologists [historian George Wise] surveyed, only a third were fulfilled”. To be precise, I wonder at the use of the word “only”. Anyone able to predict the outcome of a horse race one out of three times could make a handsome income. It seems an impressive success rate to me.

Kenneth D. Pimble

Poynter Center for the Study of Ethics and American Institutions, Indiana University, Bloomington, Indiana 47405-3602, USA

BOOKS & ARTS

Evolution's highest branches

An intriguing tour around and, especially, up the tree of life.



Creatures of Accident: The Rise of the Animal Kingdom

by Wallace Arthur

Hill & Wang: 2006. 243 pp. \$25

Matthew A. Wills

All life on Earth, no matter how complex, shares a common, very simple, ancestor. You and the bacteria in your gut have been evolving away from this starting point for precisely the same length of time. So how is it that you are capable of perusing a magazine, whereas your gut bacteria look little different from the most ancient prokaryote fossils we know? The answer, in a nutshell, is development.

From a starting point of maximum simplicity, there are only two directions you can take. One, to borrow an analogy provided by Wallace Arthur in his book *Creatures of Accident*, is 'sideways' — to diversify without increasing in complexity, an endless fugue on the bacterial theme. Undeniably, this strategy has worked very well: there are more prokaryotes on the planet today than anything else. The other direction is 'up', towards more complex and elaborate forms. But this is risky and difficult, witnessed by the gap of perhaps three billion years between the first prokaryotes and the first

animals. Assembling a complex body means you have to glue cells together, organize and differentiate them. Dividing one adult into two simply won't do; you need to develop.

Evolution seems to have produced developmental complexity very differently from the way an engineer would have designed it. Components, from genes to legs, are duplicated, and sister copies freed up and recruited elsewhere. Biologists in the 1990s were astonished to discover that flies and humans share the same underlying *Hox* gene mechanism to pattern their bodies from head to tail. But they were even more astonished to find that part of the same mechanism is cannibalized to regulate the development of vertebrate limbs.

So, how should we visualize the tree of life at the broadest scale? Arthur considers all the popular devices: the ladder (a one-dimensional, anthropocentric model, with its connotations of progress); the cladist's dendrogram (in some respects the opposite, egalitarian view, which places all species at the tips and ignores time); and the geologist's fossil-range chart (which considers time alone). Only by combining elements of all these can we get something approaching a meaningful picture.

But the story of increasing complexity —

the 'up' — has been downplayed to an absurd degree, and Arthur's book seeks to remedy this. The complexity of any given lineage can decrease as well as increase, but the 'average' and 'maximum' complexity (in a statistical sense, mass extinctions aside) have increased right up to recent times. The human brain enables behaviours vastly more elaborate than those of even our closest relatives. But is this an inevitable corollary of evolution? Arthur — like Simon Conway Morris in his excellent book *Life's Solution* (Cambridge University Press, 2003) — thinks it probably is.

The subtext and closing chapters are an attack on fundamentalism of every stripe. Pleasingly, there is little explicit tilting at the straw men of creationism and intelligent design: atheism and theism are equally irrational in Arthur's view. Richard Dawkins, famously, has no truck with this stance "because the same could be said of Father Christmas and tooth fairies". But to lump these with the 'God hypothesis' is facile: the latter has at least been entertained by some celebrated evolutionary thinkers (George Gaylord Simpson, Theodosius Dobzhansky, Stephen Jay Gould and even Charles Darwin from time to time). Dawkins is glib on this: if only we would look at the evidence, God would disappear in a puff of logic. Arthur gives the lie to this, finding nothing to evaluate: "I see no more reason for a rational scientist to be a committed atheist than to be a committed theist."

Arthur's science, although beautifully simplified and readable, is never simplistic. Unfortunately, theological arguments lose out: a liberal, humanist world-view doesn't inevitably flow from the story in the preceding chapters (or, if it does, I needed to hear the arguments hammered out), and I hit the glossary with a bump. But this is a small criticism. Readers whose interest is piqued should check out *Dawkins' God* by Alister McGrath (Blackwell, 2004) and *Pascal's Fire* by Keith Ward (Oneworld, 2006).

Creatures of Accident richly deserves to be read widely. With biology teaching becoming increasingly reductionist, it offers the (gloriously coined) "megaevolutionary" perspective. Lay readers will find plenty of surprises, and those with a more scientific background will find Arthur's straightforward and non-posturing style hugely enjoyable. ■

Matthew A. Wills is a lecturer in evolutionary biology at the University of Bath, Claverton Down, Bath BA2 7AY, UK.

JAPACK/PHOTOLIBRARY.COM

Psychoneuroendocrinology

The Female Brain

by Louann Brizendine

Morgan Road: 2006. 279 pp. \$24.95

Rebecca M. Young & Evan Balaban

In an age when one in three American adults firmly rejects evolution as false, it is a daunting challenge to write a popular and accessible account of the endocrinology, pharmacology, neurobiology, development and evolution of human sex differences. As a result, we are inclined to give authors who take up the challenge a certain amount of freedom to spruce up the facts in order to attract lay readers, who may not have the patience for the usual cautious, scientific approach.

In her book *The Female Brain*, Louann Brizendine adopts a mix of self-help, sex-specific medicine and populist neuroscience. The book advances a particularly stark version of the theory that exposure to prenatal hormones 'hard-wire' male and female brains for sex-differentiated patterns of emotion and cognition throughout life. Brizendine — director of the Women's Mood and Hormone Clinic at the University of California, San Francisco, with diplomas in neurobiology from the University of California, Berkeley, medicine from Yale University, and psychiatric training at Harvard Medical School — uses personal stories from her patients, her friends and her own life to anchor the discussion of sex differences in behaviour, hormones and the brain. The stories are the best part of the book, and it is through these that Brizendine emerges as a dedicated and sympathetic clinician. Readers whose eyes glaze over when they encounter scientific concepts will surely be drawn in.

Yet, despite the author's extensive academic credentials, *The Female Brain* disappointingly fails to meet even the most basic standards of scientific accuracy and balance. The book is riddled with scientific errors and is misleading about the processes of brain development, the neuroendocrine system, and the nature of sex differences in general. At the 'big picture' level, three errors stand out. First, human sex differences are elevated almost to the point of creating different species, yet virtually all differences in brain structure, and most differences in behaviour, are characterized by small average differences and a great deal of male-female overlap at the individual level. Second, data on structural and functional differences

in the brain are routinely framed as if they must precede all sex differences in behaviour. Finally, the focus on hormone levels to the virtual exclusion of the systems that interpret them (and the mutual regulatory interactions between receptor and secretion systems) is especially lamentable, given the book's clinical emphasis on hormone therapies.

Misrepresentations of scientific details are legion. Readers who studied biology in high school may puzzle over the invocations of the male brain with its single "dose of X



Two species? Cary Grant and Katharine Hepburn in *The Philadelphia Story*.

chromosome (there are two Xs in a girl)": is the author suggesting that X-chromosome dosage compensation is absent from female brains? Is it an improvement to dispel the myth that testosterone is a "male hormone" only to call it the "sex and aggression hormone"? (If each hormone needs a sound bite, "confidence and sense of well-being hormone" might better fit the data.) Ironically, at the intracellular level, much of the differentiation of the "testosterone-formed male brain" is accomplished by oestrogens. Fostering such misleading metaphors may prevent broader understanding.

The text is rife with 'facts' that do not exist in the supporting references. A typical example is the claim that young boys "physically cannot

hear" the cues in the intonation of adult human female voices that girls can, "just as bats can hear sounds that even cats and dogs cannot". The references provided (including a paper on songbird brains) require major misunderstanding or misrepresentation to be twisted into such a statement, a state of affairs that is repeated throughout the book.

Like other popular books on the biology of human nature, *The Female Brain* has a rigid plot line: the foil of 'political correctness' against which the author wages a struggle for truth. We are told that the media, feminists, pointy-headed intellectuals and a vaguely specified 'culture' dogmatically insist that gender or racial differences in personality and behaviour are entirely cultural, an observation that is hard to reconcile with the volume and tone of media attention to the biology of gender and sexuality.

Such assertions require empirical support. This genre loves to dwell on childhood toy preferences: little girls cradle inanimate, 'boy-coded' objects as if they were baby dolls (here, as is often the case, it's a fire engine); and little boys turn harmless objects into weapons (our favourite is the boy who bites his toast into a gun in Deborah Blum's *Sex on the Brain* (Allen Lane, 1997)). The emphasis on myth-busting turns into a vehicle for dressing the myth up in new clothes — such as Simon Baron-Cohen's recent hypothesis that the 'male brain' is hard-wired for 'systematizing', and the 'female brain' is hard-wired for 'empathizing' — there is no shortage of pseudo-scientific ways of saying 'thinkers' and 'feelers'. The problem with such explanations of sex differences is not that they are overly biological, but that they are fundamentally non-biological and explain nothing.

Ultimately, this book, like others in its genre, is a melodrama. Common beliefs are recast as imperilled and then saved. Stark, predictable protagonists (an initial "cast of neuro-hormone characters" that reads like a guide to astrological signs) interact linearly with foreseeable results. The melodrama obscures how biology matters; neither hormones nor brains are pink or blue. Our attempts to understand the biology of human behaviour cannot move forward until we try to explain things as they are, not as we would like them to be. ■

Rebecca M. Young is at the Women's Studies Department, Barnard College, Columbia University, New York, New York 10027, USA.

Evan Balaban is in the Behavioral Neurosciences Program, McGill University, Montreal QC H3A 1B1, Canada.

A comprehensive medical history

The Western Medical Tradition: 1800 to 2000

by W. F. Bynum, Anne Hardy, Stephen Jacyna, Christopher Lawrence & E. M. (Tilli) Tansey

Cambridge University Press: 2006. 614 pp. \$90, £50 (hbk); £19.99, \$29.99 (pbk). Available as a two-volume set with *The Western Medical Tradition: 800 BC to AD 1800*. \$70, £40

Andreas-Holger Maehle

"The social history of medicine has come of age," stated historian Andrew Wear nearly 15 years ago in the introduction to *Medicine in Society* (Cambridge University Press, 1992), a collection of historical essays that became popular as a textbook. After reading *The Western Medical Tradition: 1800 to 2000*, by Bill Bynum and colleagues at the Wellcome Trust Centre for the History of Medicine at University College London, my diagnosis is that the subject has now fully matured.

Conceptualized as a companion volume to *The Western Medical Tradition: 800 BC to AD 1800* by authors from the same centre, the new book provides in four large chapters a comprehensive synthesis of the development of medicine in the context of nineteenth- and twentieth-century Western society. Two general themes emerge: one is the increasing effect of the natural sciences in shaping medicine against the background of industrialization and modernization; the other is the gradual acceptance, by governments and the public alike, of

Western orthodox medicine as the arbiter in matters of individual and collective health. The history of 'alternative' medicine, although not ignored, does not receive particular treatment in this otherwise wide-ranging historical survey. There is, for example, no sustained discussion of the history of homeopathy.

Inevitably, perhaps, the broad overview in this book reflects certain trends and focal points in the historiography of medicine over the past 20 years or so. Accordingly, the history of public health and social welfare, of epidemics, of medicine in wartime, and the rise of modern hospital medicine take up a lot of space. Scientists central to the history of medicine, such as Rudolf Virchow, Louis Pasteur and Robert Koch, are treated in due detail. The book also raises awareness of historiographic methods. Stephen Jacyna, for example, in his chapter on the first half of the nineteenth century, discusses the approaches of Marxism and Michel Foucault to the history of medicine.

The book has a strong international perspective, covering developments in Germany and France, as well as Britain and the United States. It was not written strictly as a comparative history, but still provides valuable insight into different national styles of medicine. For instance, the strong drive towards medical specialization in Germany, the United States and France is contrasted with a resistance to this trend by British doctors that lasted until after the First World War. Medicine in Nazi Germany is treated in a separate section, but the presence of eugenic ideas and practices in

Western countries during the 1920s and '30s is emphasized in other parts of the book. Given this international approach, it is regrettable that the otherwise useful bibliographies for each chapter cite only English-language publications. This may be appropriate for a volume that is bound to become a key textbook for UK and US courses on the history of medicine, but the book's reach could have been made even wider by including, say, some French- and German-language scholarship in the field.

Another merit of this volume is that it contains one of the very few summaries of the history of Western medicine after 1945. The authors of this chapter, Anne Hardy and Tilli Tansey, focus on Anglo-American developments, but regularly contextualize them with data from various European countries. They also address the apparent paradox that despite undeniable progress in diagnostics, therapy and disease prevention, public perception of medicine and biomedical research has become increasingly critical since the 1970s. The story of progress in medicine is also qualified by public criticism of the inequality in health-care provision between developing and developed countries, especially in the face of the HIV pandemic, the resurgence of tuberculosis and continuing problems with malaria control.

As this book shows, traditions in medicine and science can have a powerful influence on how health care is shaped for the future. I therefore recommend the book to anyone with a serious interest in understanding the interactions between the social and scientific events that have formed modern medicine. ■

Andreas-Holger Maehle is head of the Department of Philosophy, Durham University, Durham DH1 3HN, UK.

The background buzz

Noise

by Bart Kosko

Viking: 2006. 272 pp. \$24.95

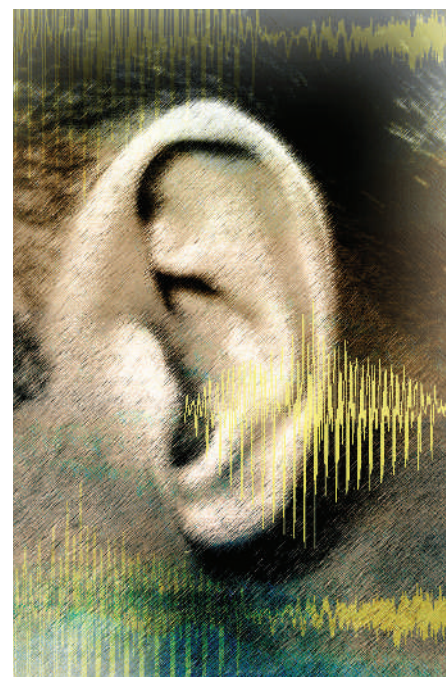
P. V. E. McClintock

Every macroscopic system is subject to noise, from the random fluctuations of its environment as well as its own internal fluctuations. If, like Bart Kosko, you consider things such as e-mail spam and extraneous signals and objects to be noise, then your immersion seems even more overwhelming. Noise is not a trivial detail, and in general it cannot just be ignored on the grounds that it will somehow average out. In the case of nonlinear systems, far from averaging out, it gives rise to a diverse range of important phenomena, such as stochastic resonance.

In *Noise*, Kosko takes the topic of noise, using the widest possible definition, and tries to make it accessible and interesting to the general reader. He opens with his central

thesis that noise is "an unwanted signal" and develops ideas from information theory of noise in digital signals. The next two chapters cover noise in the everyday sense of unwanted acoustic disturbances. There is anecdotal material about unneighbourly behaviour, litigation, crematoria, noise from aircraft, the law on trespass, damage to hearing, noise-induced stress, and undersea noise involving the US Navy, dolphins and whales.

Much of the hard science, and its exposition in simple terms, is consigned to the (long) fourth chapter. Kosko covers a remarkably wide range of topics, from cosmology and photosynthesis to white and coloured noise, and the central-limit theorem. This is followed by a chapter including Fourier techniques and spread-spectrum encryption. The book ends with a discussion of stochastic resonance, an interesting phenomenon in which noise plays a creative role, rather than its usual destructive one. Here the addition of noise to a nonlinear



HANNAH GAL/SPL

In one ear: there's more to noise than unwanted sound, and it's not all bad.

system can cause amplification and enhance the signal-to-noise ratio of a signal passing through the system.

Each chapter is prefaced with ten or so intriguing and provocative quotations that occupy several pages. There are also more than 60 pages of notes at the end of the book, arranged by chapter, expanding on the material in the main text. Many of them are useful and interesting, but in the end I got tired of continually turning pages to find them; they would have been far more convenient as footnotes.

Has Kosko succeeded in his aims? I believe he has, to a large extent, although in one or two places he illustrates the G. K. Chesterton dictum that "He who simplifies simply lies" by conveying a misleading impression. For example, few readers will appreciate that stochastic

resonance can do nothing to enhance the signal-to-noise ratio of a given signal. It can certainly ameliorate the information loss that otherwise occurs when a signal passes through a nonlinear system, but I very much doubt whether readers will understand this fact from Kosko's discussion. I suspect, though, that schoolboy howlers such as "The brain consumes about 20 watts of power each day" are not from the author's pen.

I am surprised that Kosko omits all mention of optimal fluctuational paths, given their conceptual simplicity and the nice way they link together many of the other ideas he presents. Most of the interesting and important events in noisy systems involve such optimal paths, including chemical reactions, mutations in genes, and failures of lasers and electronic

devices. In all these cases, the system fluctuates near an attractor of some kind for a long time, and then travels along an optimal path to a different attractor. Remarkably, despite the noisy driving force, these paths are deterministic in character.

While accepting the author's broad view of what constitutes noise, I also feel that some of his writing introduces its own noise, through the unnecessary introduction of a multiplicity of distracting ideas that are tangential (or irrelevant) to the point being discussed. What is certain, however, is that every reader of this book will end up learning something new and interesting.

P. V. E. McClintock is in the Department of Physics, Lancaster University, Lancaster LA1 4YB, UK.

A modern megalith

Mariko Mori's glass sculpture responds to the death of stars.

Martin Kemp

Prehistoric standing stones and rings, many erected more than 5,000 years ago, are awesome achievements. Not surprisingly, the greatest of them, such as Stonehenge, have served over the centuries as magnets for legends and mystical mumbo-jumbo.

The reality being revealed by modern archaeoastronomy provokes almost as much wonder as the legends. Perhaps most astonishing are the astronomical alignments that have been demonstrated in megaliths. It is clear that our 'primitive' forebears not only observed celestial phenomena with remarkable precision, but also built their great stone monuments as a means of relating their earthly existence to cosmological events far beyond their reach.

Such astronomical megaliths serve as the inspiration for an installation by that most high-tech of artists, Mariko Mori. A student of fashion in her native Japan, she worked briefly as a model before studying art in London and now lives in New York. She began by exploiting multimedia to fashion herself into a futuristic 'cyber-chick', transformed into a synthetic fantasy of kitsch sexuality, far removed from the ragged desires of our organic reality.

Superficially — and it is easy to see such work as superficial — she seemed to belong to a late species of pop art, delighting ironically in the sheen of slick popular imagery. However, she has insisted on a more serious purpose, adducing her immersion in Buddhist philosophy to stress the interconnectedness of all things, via art, science and technology.

With her earlier work it was difficult not to see these high claims as somewhat forced. However, her recent creations have laid this



Tom Na H-iu: the output of a neutrino detector becomes a meditation on the soul.

problem to rest. She has been collaborating with scientists to produce experiences that are startling in their technical sophistication, yet evoke both the inner world of our minds and the outer worlds of the cosmos.

At the 2005 Venice Biennale she exhibited *Wave UFO*, a futuristic pod in which the brainwave data from electrodes attached to three participants were projected on the ceiling as mutating coloured shapes. Three kinds of waves — alpha (blue), beta (pink) and theta (yellow) — were used to render visible endlessly variable arrays of mental processes. Now, at the 2006 Singapore Biennale, she is showing *Tom Na H-iu*, a

3-metre-high radiant glass monolith that is plugged into cosmic radiation.

The translucent megalith is suffused by light from an internal LED, controlled by a computer, which is, in turn, linked to the Super-Kamiokande detector used in the Kamioka Observatory in Japan to detect neutrinos from outer space. Among the neutrinos that govern the megalith's light emissions are some that emanate from exploding, dying stars — supernovae. Programmed to respond to the detection of different neutrinos in real time, the sculpture glows with colour-coded traces of ancient violence from remote regions of the cosmos.

In Celtic mythology, say the work's promoters, *Tom Na H-iu* is a place where the souls of the dead linger before being reborn, and the Celtic standing stones that inspired the artwork were believed to play a role in this spiritual transmigration. Mori is tapping into the puzzles of birth and death across enormous distances and deep time. The microscopic is a mirror of the cosmic. The lives of our minds and bodies are integral to the processes that govern the Universe.

Just as our Neolithic ancestors reached out to the stars, tuning into the apparently eternal patterns that ruled their seasonal lives, so Mori's sculpture resonates to the beat of celestial time. However, that beat is no longer that of the traditional music of the orderly spheres, but drums out the death march of great stars.

A larger version of *Tom Na H-iu* is on show with other work by Mori at the London gallery Albion until 22 December.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK. His new book, *Seen | Unseen*, is published by Oxford University Press.

SHIRAIISHI CONTEMPORARY ART, TOKYO

PARTICLE PHYSICS

Did the Big Bang boil?

Frank Wilczek

Standard theories tell us that, at some point in the Universe's evolution, free quarks and gluons must have become bound together into the hadronic matter we see today. But was this transition abrupt or smooth?

The idea of phase transitions — abrupt changes in the state of matter — is familiar from such common sights as the bubbling water in a boiling kettle. Phase transitions on a grand scale may have taken place in the early Universe, both enriching and complicating Big Bang cosmology. For example, the early Universe's gas of quasi-free quarks and gluons must at some point have condensed into composite particles bound together by the nuclear strong force. These particles, made up of quarks stuck together by gluons, which act as the mediator particles of the strong force, are the baryons — the class of particle that includes protons and neutrons — and the mesons of today's normal 'hadronic' matter.

It had once seemed plausible that this phase change was marked by a true, abrupt thermodynamic phase transition, such as that seen in the kettle, at a temperature of around 10^{13} kelvin. Through a difficult and fundamental calculation, set out on page 675 of this issue, Aoki *et al.*¹ have demonstrated that this change, although drastic, actually sets in smoothly as the temperature falls. Their methods could be used to illuminate several other open questions.

According to conventional Big Bang cosmology, the early Universe contained matter in thermal equilibrium at extremely high temperatures. As the Universe expanded and cooled, neutrinos and eventually photons ceased to interact significantly, and fell out of equilibrium. These freeze-outs occurred at temperatures of 10^{10} K and 10^3 K, respectively. (The relic photons, redshifted down to a temperature of just around 3 K, owing to the Universe's subsequent expansion, constitute the present-day cosmic microwave background radiation.) But at the earliest times, before neutrinos and photons decoupled from the rest of matter, densities and energies were so high, and reaction rates therefore so fast, that any local deviations from equilibrium were quickly repaired. That might seem to be a recipe for dullness. But as anyone who's seen the agitation in a boiling kettle will know, phase transitions can spice up the action.

The most dramatic consequences come by way of first-order phase transitions, which involve discontinuous changes in the organization of matter. Suppose phase A has a lower free

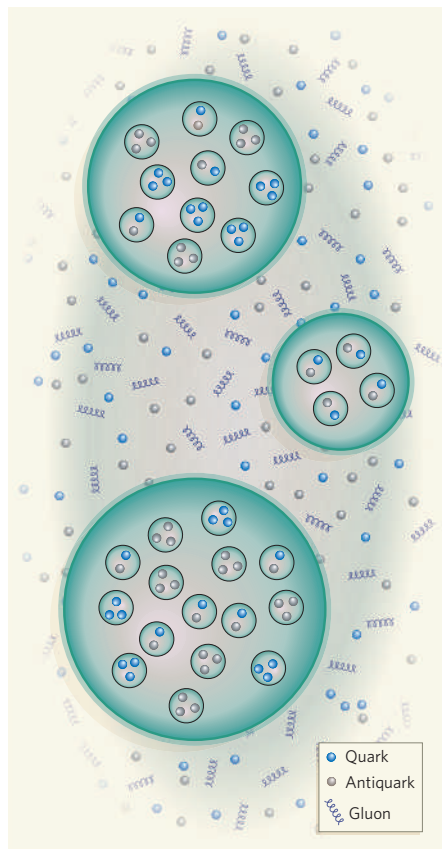


Figure 1 | Entering a new phase. A first-order transition from a quark–gluon plasma to normal hadronic matter might look like this, with bubbles of the new phase forming in the old. Outside the bubbles, quarks (blue), antiquarks (grey) and gluons (squiggles) roam free. Inside the bubbles, they are bound into mesons (quark–antiquark pairings), baryons (three quarks) and antibaryons (three antiquarks).

energy above some critical temperature T^* , and that below T^* phase B has a lower free energy. The phase with the lower free energy is always the natural choice, so as we start from phase A at high temperature and cool through T^* , bubbles of B may appear. At temperatures below T^* , a sufficiently large bubble of phase B can grow, as gains in volume free energy within the bubble overcome surface tension at the interface. Eventually, then, we will have only B.

It may, however, take a very long time to nucleate a sufficiently large bubble. In the meantime, A persists below T^* : this is the phenomenon of supercooling. In a cosmological context, the unusual equation of state associated with supercooling can induce a period of faster-than-light expansion — the phenomenon known as inflation. Alternatively, the bubbles might suddenly appear in many different places, then grow and collide. Those violent events can leave their imprint on the gravitational field, leading to seed fluctuations from which galaxies or black holes might eventually evolve. Gravitational waves might be yet another result, awaiting detection today as a relic background. Different processing of matter within the regions of A and B might also lead to deviations from standard predictions for the abundances of the light elements, or even package matter into unusual forms, such as superdense quark matter.

At extremely high temperatures, above 10^{13} K, quantum chromodynamics (QCD), the theory of the strong nuclear interaction, predicts that matter can be accurately described as a weakly interacting gas of quarks and gluons. That seems very different from what we now observe, where quarks and gluons occur only as hidden constituents inside baryons and mesons.

There is also another much less obvious, but no less profound, difference. At low temperatures, the world is filled with a condensate of quark–antiquark pairs. Our senses are designed, quite sensibly, to overlook that featureless and unchanging background, but it has been detected through its subtle but unmistakable influence on the behaviour of particles we do see. For example, it renders the lightest mesons, the π -mesons, much lighter than they would otherwise be. QCD also predicts that, at high temperatures in the early Universe, the quark–antiquark condensate would have disappeared.

Qualitatively described, these changes in the constitution of matter seem quite drastic. Thus, it seemed plausible to many theorists that the conversion from high-temperature to low-temperature QCD would have to be accompanied by a first-order phase transition (Fig. 1). That

intuition was never conclusive: the distances over which quarks roam might just get smaller and smaller, and the condensate less and less tenuous, with no abrupt jumps in the Universe's constitution. But it is hard to visualize appropriate intermediate states.

Fortunately, the equations of QCD are precisely defined and reliable, so the question can be settled by calculation. Unfortunately, the equations are hard to solve, so until now a definitive answer has been elusive. By exploiting the full power of modern, large-scale computing, Aoki *et al.*¹ have answered the question, 'Did the Big Bang boil?'. The answer, as far as the quark-hadron transition is concerned, is 'No'. QCD evolves smoothly with temperature; there is no thermodynamic phase transition.

Thus, Aoki and collaborators have brought closure to some intriguing speculations, and reinforced the foundations of Big Bang cosmology. For the powerful techniques they used, however, it is only the beginning.

According to a famous Berkeley graffito, "Reality is a crutch." Although QCD with the values of quark masses that we have in our Universe apparently does not feature a true phase transition, it will be important to explore whether that conclusion remains true if those values are changed. There are quite firm and detailed theoretical predictions that phase transitions of both first and second order will occur in this case. (Second-order transitions are continuous, but feature more subtle singularities. They, too, can lead to large-scale density fluctuations.)

Also, collisions of heavy ions at almost the speed of light are being studied at the Relativistic Heavy Ion Collider at the Brookhaven National Laboratory on Long Island, and will soon be studied at CERN's Large Hadron Collider near Geneva. Such collisions can produce fireballs with a significant excess of quarks over antiquarks, or different effective temperatures for quarks and gluons — possibilities that did not occur in the cosmic Big Bang. In those new circumstances, do true phase transitions occur? Once we learn the answers to such questions, a dramatic confrontation between theory and experiment might be staged.

Looking further ahead, we will progressively learn more about so-called electroweak symmetry breaking (the question of how the mediator particles of the electromagnetic and weak forces came to have different masses), the possible unification of all the fundamental forces, and other physics beyond the standard model. Then, the question of possible phase transitions other than that considered by Aoki *et al.*¹, occurring at still higher temperatures, will come more sharply into focus. With luck, we will learn enough to compute definitive answers. In the meantime, we can have fun developing the necessary intellectual strength through exercise on QCD and its variants. ■ Frank Wilczek is in the Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. e-mail: wilczek@mit.edu

1. Aoki, Y., Endrődi, G., Fodor, Z., Katz, S. D. & Szabó, K. K. *Nature* **443**, 675–678 (2006).

EVOLUTIONARY BIOLOGY

Fly eyes get the whole picture

Kevin Moses

The compound eyes of ancestral flies picked up only one picture point in each facet. The evolution of a means to split up the light-sensitive cells increased this number to seven, boosting the eye's resolution greatly.

"As anyone knows who has tried to catch one, flies see extremely well."

— Ian A. Meinertzhagen

Clearly, from the fly's perspective, seeing well is a very good thing. Those ancestral flies that did not see well contributed more to the evolution of frogs than to the fly lineage. On page 696 of this issue, Zelhof and colleagues¹ reveal how flies improved the resolution of their eyes dramatically by evolving a specific modification to how their photoreceptor cells stick together*.

From fossils (mostly trilobites), we know that arthropods have used faceted, compound eyes ever since the Cambrian era, about 540

million years ago. Compound eyes produce a mosaic view of the environment because each of the eye facets (or ommatidia) is directed at a slightly different point in space. The resolution of the image is limited by the number of ommatidia and the angles between them — to get more picture points requires more ommatidia. Even though insect photoreceptor cells have evolved to be very small, there is a fitness cost to growing a bigger eye. So, for most of the past 540 million years, arthropods have had to deal with the problem that better vision demands a bigger eye.

But there are different vision requirements at night from those for the daytime, so the optimal balance in the trade-off between resolution and size varies. We humans solve this conundrum by

having two eyes built into each of ours. About 95% of our retina is covered in rod cells, which cannot distinguish colours but work well at night. For daylight vision, in the centre of our retina (the macula) about 5% of our photoreceptor cells (the cones) can distinguish colour but are less sensitive. Furthermore, our eyes are wired for low resolution at night, when it is more important to see every photon, and for high resolution in daylight. We achieve this by having many rod cells stimulate one of the interneurons that process the retinal signal and pass it on to the brain, whereas in the centre of the retina each cone cell feeds only one interneuron.

Insects solve the day–night trade-off in a different way². Insects that must see well in the daytime have 'apposition' eyes, with seven or eight photoreceptor cells clustered under each ommatidial lens. These cells are arranged so that their light-gathering surfaces (the rhabdomeres) are fused at the facet's centre to make a structure called the rhabdom, so that they all receive light along the central optical axis (Fig. 1a). The entire facet is sheathed in a tube of pigment cells so that light coming in at other angles is removed (Fig. 1b). So most photons are lost, but resolution is improved. Night-flying insects have 'superposition' eyes, which lack most of the sheathing pigment so that light from a wider field can act on each photoreceptor (Fig. 1c). This is more sensitive but sacrifices resolution. Many insects are able to shift pigment within their pigment cells so that their eye can function in both ways: in effect, they raise their curtains at night.

About 100 million years ago, the diptera (mosquitoes and flies) developed a clever trick to squeeze more picture points out of their apposition eyes — they separated the photoreceptor rhabdomeres in each ommatidium so that, instead of seeing one picture point per facet, they now see seven (to the dismay of frogs ever since; Fig. 1d). This has been called a 'neural superposition' eye³ and also involves some equally clever rewiring in the fly's brain⁴ (Fig. 1e). In their paper¹, Zelhof and colleagues describe how diptera managed to develop this system.

Zelhof *et al.* looked for fruitfly mutants that have no inter-rhabdomeral space but have the rhabdomeres collapsed back into the centre of the facet like lower insects. They found loss-of-function mutations in two genes that have this effect, but otherwise leave the rhabdomeres intact. One gene they name *spacemaker*, and the other is a gene that encodes a previously known protein called Prominin that helps to form cell–cell contacts. The Spacemaker product was predicted to be a secreted protein, and Zelhof and colleagues show that it is indeed normally present in the inter-rhabdomeral space, beginning in the pupal stage when the rhabdomeres are formed. Furthermore, they show that Spacemaker protein can bind to Prominin (in tissue-culture binding and cell-adhesion assays). Their data suggest that Spacemaker and Prominin normally function to oppose a third, previously

*This article and the paper concerned¹ were published online on 1 October 2006.

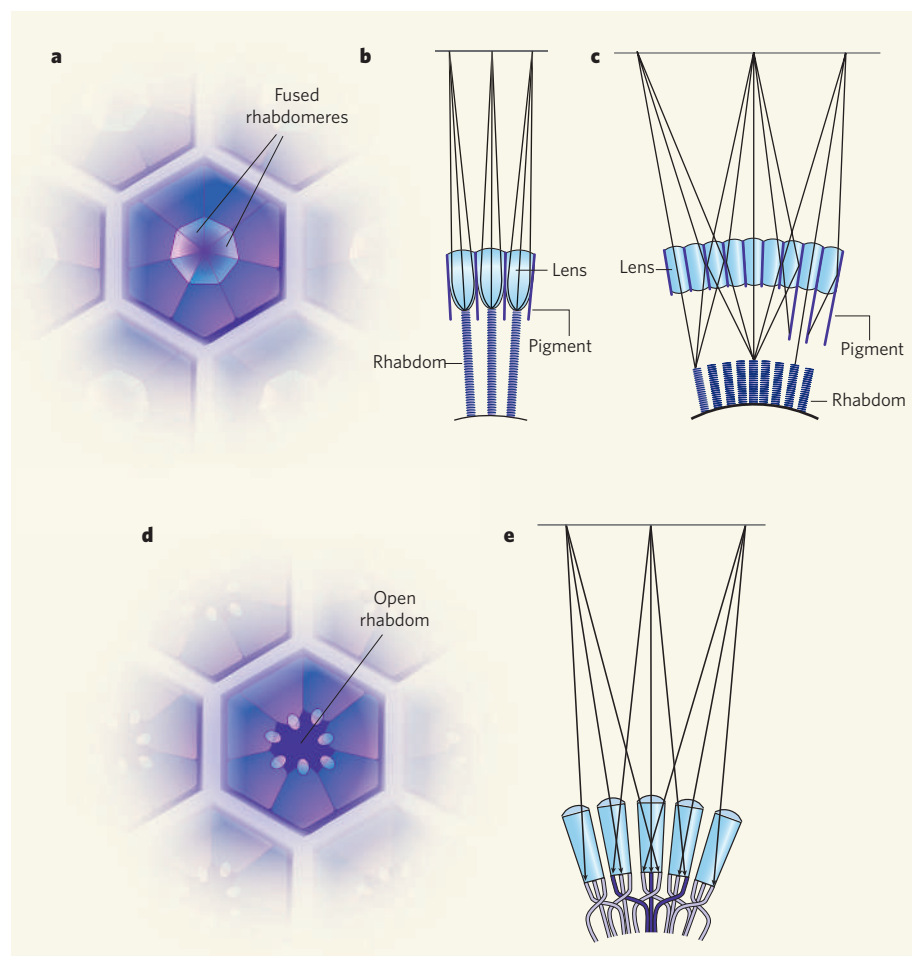


Figure 1 | How fly eyes differ from those of other arthropods. **a, d,** Surface view of a single eye facet (ommatidium) of a compound eye. **b, c, e,** Ommatidia seen as though sliced through their long axis. **a,** In the ancestral fly eye and in modern-day insects such as bees and beetles, the light-capturing surfaces (rhabdomeres) are fused in the centre of the ommatidium. **b,** In 'apposition' eyes, pigment cells sheath each ommatidium, so light is received only down the central axis and image resolution is improved. **c,** Insects that need good night vision have 'superposition' eyes, which lack most of this pigment, so light is received from all angles. In this case more photons are received, but the resolution of the image is reduced. **d,** Modern flies and some mosquitoes have open rhabdoms, so that instead of one picture point per ommatidium, they can perceive seven. This requires some rewiring of the underlying neurons so that light received from different angles in one ommatidium can be resolved (neural superposition, **e**). Zelhof *et al.*¹ find that expression of the Spacemaker protein is responsible for the opening up of the rhabdomeres.

known protein, Chaoptin⁵, to divide the upper part of the photoreceptor cell membrane into rhabdomere and 'stalk' domains. These proteins also delicately regulate cell-cell contacts so that the rhabdoms stay intact, but the adjacent rhabdomeres don't stick to each other.

Zelhof *et al.* find that Spacemaker is absent from the eyes of insects with fused rhabdom systems (bees and beetles), but is present in other diptera with open rhabdoms (the house fly and a mosquito). Prominin and Chaoptin are present in all cases, and so it seems that Spacemaker is the crucial factor in making the inter-rhabdomeral space. In a most satisfying control experiment, they targeted the expression of Spacemaker to another type of fly eye: the ocelli (simple eyes on the top of the head). The three ocelli have many photoreceptor cells, but their rhabdomeres are normally fused. When Zelhof *et al.* expressed

Spacemaker in the ocelli, they found that an inter-rhabdomeral space opens up.

These experiments suggest that the diptera may have 'opened their eyes' (invented neural superposition) by a single change: reprogramming the expression of Spacemaker for novel expression in the ommatidia. It is not often that we get such a clear glimpse of the blind watchmaker at work.

Kevin Moses is at the Howard Hughes Medical Institute, Janelia Farm Research Campus, 19700 Helix Drive, Ashburn, Virginia 20147, USA. e-mail: mosesk@hhmi.org

1. Zelhof, A. C., Hardy, R. W., Becker, A. & Zuker, C. S. *Nature* **443**, 696–699 (2006).
2. Land, M. F. & Fernald, R. D. *Annu. Rev. Neurosci.* **15**, 1–29 (1992).
3. Franceschini, N. in *Photoreceptor Optics* (eds Snyder, A. W. & Menzel, R.) 98–125 (Springer, Heidelberg, 1975).
4. Meinertzhagen, I. A. *Neuron* **28**, 310–313 (2000).
5. van Vactor, D. L. J., Krantz, D. E., Reinke, R. & Zipursky, S. L. *Cell* **52**, 281–290 (1988).



50 YEARS AGO

A new and rapid technique of characterizing the chemical properties of a protein in considerable detail has been devised; by its application a specific difference is found in the sequence of amino-acid residues of normal and sickle-cell haemoglobin. This difference appears to be confined to one small section of the polypeptide chains... The action of trypsin on proteins is at present the most reliable way of splitting a peptide chain at specific peptide bonds... Small differences in the two proteins will result in small changes in one or more of [the resulting] peptides. These should be detectable when the mixture is examined by a two-dimensional combination of paper electrophoresis and paper chromatography. It was decided to call the resulting chromatogram the 'finger print' of the protein.

V. M. Ingram

From *Nature* 13 October 1956.

100 YEARS AGO

In my address at York I urged biometricians to make sure that the problems they seek to elucidate are sound from a biological point of view. When asked by Prof. Pearson for an instance of failure in this respect I gave him, while away on my holiday, and in a private letter, Dr. Pearl's paper. He has now seen fit, although I twice asked him to wait for a full answer until my return to Cambridge, to challenge me to show in the pages of *NATURE* how my advice was applicable to that paper. I must leave your readers to judge how far I have succeeded in so doing.

The task has been far from an agreeable one. I should never have thought of singling Dr. Pearl's paper out for public criticism in this manner had I not been challenged to do so. I can only say that if he feels himself aggrieved at the result, he can be in no doubt whom he has to thank.

J. J. Lister

From *Nature* 11 October 1906.

50 & 100 YEARS AGO

PHYSICAL CHEMISTRY

Seeds of phase change

Daan Frenkel

An effective but counter-intuitive trick to obtain highly ordered protein crystals is to 'seed' particles on disordered, porous surfaces. Computer simulations provide an explanation for the success of this strategy.

The highly ordered, three-dimensional arrays of molecules that we know as crystals can form spontaneously from the disordered liquid phase. To speed up the rate of crystal formation, crystal-growers often use 'seeds' — templates that can kick-start crystallization. A few years ago, experiments on protein crystallization¹ unexpectedly showed that highly disordered, porous particles can be particularly effective in initiating the growth of perfectly ordered crystals. Writing in *Physical Review Letters*², Page and Sear now explain why the disorder of porous templates allows them to behave as universal crystallization agents.

In many areas of science and technology, it is essential to control crystallization. For example, proteins must first be crystallized if we are to determine their structures from X-ray crystallography; pharmaceutical compounds must be prepared in their optimal crystal forms to maximize their effectiveness; and careful design is needed to make photonic crystals, the periodic arrays of micrometre-sized colloidal particles that should eventually find applications in optical devices.

One of the problems facing the crystal-grower is that crystals often don't form from solution, even when that solution is highly concentrated with the compound to be crystallized. The nineteenth-century physicist J. Willard Gibbs first explained this phenomenon by pointing out that, under conditions where large crystals are stable, the microscopic crystallites — the precursors of bulk crystal formation — are unstable and should disintegrate. But every crystal must start off small, so this implies that large crystals would never form. Brownian motion comes to the rescue: because of the incessant

thermal motion of molecules in solution, small crystallites form and break up continuously, but — very occasionally — a crystallite may form that is large enough to continue growing. This process is called crystal nucleation, and the smallest crystallite that can grow to become a bulk crystal is known as the critical nucleus.

In many cases of practical interest, the rate of crystal nucleation is vanishingly small. This is often a serious problem in the preparation of protein crystals. One way to speed up crystal nucleation is to seed the solution, using particles that have the same surface structure as the desired crystal. Of course, the best seed is a small crystal of the material that one wishes to crystallize, but clearly such seeds are not available if that material has never been crystallized before. An alternative is to use other seed materials that have the right surface patterns. This approach is quite successful for preparing colloidal crystals³, but less so for protein crystals.

It was therefore a surprise to find that a variety of proteins will crystallize on seeds of microporous silicon⁴. The characteristics of this seed material are almost the opposite of what one would expect of a good crystallization template: it is sponge-like, with a highly disordered surface full of holes and crevices. The idea of crystallization on such a rough surface is counter-intuitive because, as computer simulations on colloidal crystallization have shown, the curvature of the surface makes nucleation more difficult⁵. Crystals attached to curved surfaces are deformed, and are therefore less stable than crystals on flat surfaces. Even if a stress-free crystal does form on a rounded surface, its surface cannot accurately follow the curvature of the substrate;

again, this would not favour crystallization.

Now, however, Page and Sear² suggest an explanation for the unexpected crystallization of proteins on a porous surface. They argue that, among the myriad pores at the seed's surface, there will always be some that have exactly the right shape to accommodate a small crystallite. For instance, for a cubic crystal, such a pore would have a rectangular prism shape.

The authors then address the question of what the optimum pore size is for crystal nucleation. The answer is not simple. They consider a highly simplified model in which the width of the pore is easily varied, and find that crystal nucleation proceeds in two stages: first, crystallization starts at the corners of the pore; then, once a crystal has grown to fill the pore, it breaks out to form a bulk crystal (Fig. 1). The rate at which crystals form inside the pore increases as the pore becomes narrower. But the rate at which these crystals go on to form a bulk crystal increases as the pore diameter widens. So, for narrow pores, the breakout time is rate limiting, but for wide pores, nucleation is the slowest step.

The crystallization rate is at a maximum for pore diameters somewhere between the wide and narrow extremes, and it is these optimally sized pores that dominate crystal nucleation on a microporous substrate. The great diversity of pore sizes and shapes in a microporous material increases the likelihood of an optimal pore existing on the surface of that material. In this sense, a microporous material acts as a huge, random 'library' of crystallization agents.

Understanding the microscopic mechanism behind the success of microporous nucleation agents should help us to design seeds with increased crystallizing power. This can be of considerable practical importance, as the production of high-quality protein crystals is one of the main bottlenecks in the X-ray determination of the structure of water-soluble proteins. Unfortunately, it seems unlikely that microporous seeds will help in crystallizing membrane proteins, which constitute about a third of all proteins.

Microporous materials should be particularly useful as seed substrates for protein crystallization where the aim is to make only a few crystals. Although the vast majority of pores are ineffective, this doesn't matter as long as a few work well. By analogy, we should expect that a great multitude of random substrates — or, more generally, highly diverse environments — are necessary to bring about processes where just one successful event is enough. The evolution of life may be a case in point. ■ Daan Frenkel is at the FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098SJ Amsterdam, The Netherlands. e-mail: frenkel@amolf.nl

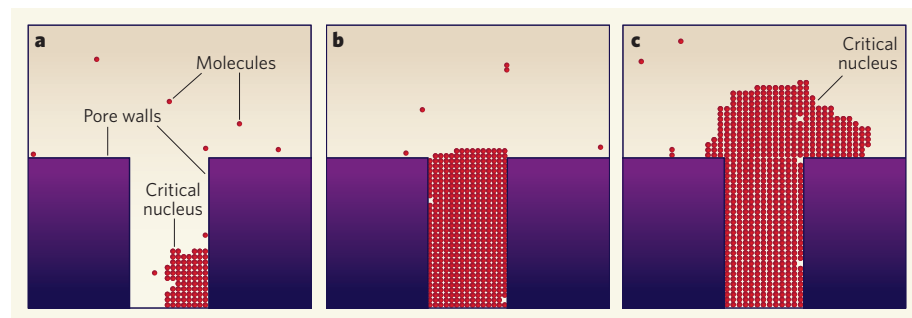


Figure 1 | Crystal nucleation in a pore. Page and Sear² propose a model to explain why porous materials induce protein crystallization. **a**, The pore walls weakly attract the molecules to be crystallized. Crystals start to grow in the corners of the pore when a critical nucleus of molecules forms. **b**, The crystal grows until the pore is filled. **c**, Another critical nucleus of molecules forms on the outside of the pore, which allows the 'breakout' of a bulk crystal from the pore. The rates of nucleus formation and breakout vary independently according to the width of the pore.

1. Chayen, N. E. *et al.* *J. Mol. Biol.* **312**, 591–595 (2001).
2. Page, A. J. & Sear, R. P. *Phys. Rev. Lett.* **97**, 065701 (2006).
3. van Blaaderen, A. *et al.* *Nature* **385**, 321–324 (1997).
4. Cacciuto, A., Auer, S. & Frenkel, D. *Nature* **428**, 404–405 (2004).

NEUROSCIENCE

Controlled capillaries

Brian A. MacVicar and Michael W. Salter

The finest scale of blood flow through the brain occurs in capillaries. Suspicions that capillary flow is regulated by cells that put the squeeze on these vessels are now borne out by detailed experiments.

The control of brain blood flow poses an intriguing 'plumbing' problem. On the one hand, high overall flows are required to maintain healthy brain function, because in humans the brain accounts for 20% of the body's energy consumption even though it forms only 5% of the total weight. On the other hand, there is a need to precisely regulate increases and decreases in flow to match the changing metabolic needs of specific brain regions. It has been known for more than 100 years¹ that the brain regulates its own blood supply. It can increase blood flow specifically to discrete regions to follow increases in neuronal activity. This principle has been exploited in functional magnetic resonance imaging and positron emission tomography, which have been extensively used to map the brain regions that are associated with different tasks².

However, the cellular mechanisms by which specific brain regions regulate blood flow, and therefore their own nutrient intake and waste removal, are not fully understood. In their paper on page 700 of this issue³, Peppiatt *et al.* implicate new players in regulating blood flow at the smallest level in the brain — cells known as pericytes.

The large blood vessels supplying the brain

are the carotid and vertebral arteries, which then branch to form the network of pial arteries covering the surface of the brain. In the cerebral cortex, the pial vessels branch into smaller arteries, which enter the brain tissue itself and are called the penetrating arterioles. These arterioles branch into secondary and tertiary arterioles, until they reach the smallest vessel supplying the brain tissue, the capillary, which is only wide enough for one red blood cell to pass through it at a time. The capillaries then feed into the venules and veins, which carry the blood away.

By virtue of the smooth muscle that surrounds them, arteries and arterioles can regulate blood flow. Various processes, including the release of mediators from the endothelial cells that line these vessels, cause contraction or relaxation of the smooth muscle cells, thereby decreasing or increasing the diameter of the artery or arteriole — *et voilà*, as we know from Poiseuille's law of fluid dynamics, blood flow can be controlled.

It had been suspected that blood flow through capillaries is also regulated, even though there are no surrounding smooth muscle cells to constrict them. The suspicions were based on the observation that capillaries

are enwrapped at intervals by a little-studied cell type with contractile capabilities, the pericyte. Peppiatt *et al.*³ now show definitively that pericytes in living tissue of the central nervous system can constrict and relax, correspondingly changing capillary diameter, and that they do this in response to changes in neuronal activity. Pericytes were identified by staining with a specific marker. The cells are spaced at intervals along capillaries, and also occur at capillary junctions. Their cell bodies abut the capillaries, with their processes enwrapping them (see Fig. 1a of the paper³ on page 701).

Peppiatt *et al.* obtained the proof that pericytes modify capillary diameter by stimulating the cells directly with microelectrodes in rat brain slices and in retinal explants. They also found that drugs that activate purinergic receptors, a type of receptor found in the surface membrane of pericytes, stimulated pericyte-induced contractions. The constriction of the capillary occurred only in discrete regions or at capillary branch points. These types of localized change in capillary diameter should be very effective in blocking the entry of red blood cells, and thus slowing blood flow. The branch-point constriction might also divert flow into a capillary supplying an active region of neurons.

Inhibiting pericyte constrictions, by lowering the levels of external calcium (a common agent of cell signalling) or applying the neurotransmitter glutamate, caused dilation of pre-constricted capillaries, indicating that pericytes exert bidirectional control over capillary diameter. Significantly, the pericytes became constricted in response to stimulation by local neurons, showing that pericytes are sensitive to changes in neuronal activity.

DEVELOPMENTAL BIOLOGY

A change of heart

Vertebrate hearts have at least two chambers. But how did these evolve from the single-chambered pumps seen in simpler organisms? While examining the development of the heart of the sea squirt, Brad Davidson and colleagues (*Genes Dev.* **20**, 2728–2738; 2006) may have chanced on the answer.

The adult sea squirt (*Ciona intestinalis*, pictured) is a classic squishy invertebrate, but as a larval tadpole it resembles a fish embryo. This puts it closer to humans on the evolutionary scale than other genetic model organisms such as fruitflies and worms, and makes it a useful system for studying certain developmental processes.

Davidson *et al.* followed the single-chambered sea-squirt heart as it

grew from two cells in the early embryo. These cells divide and specialize to form muscle cells in either the heart or the tadpole tail. The authors find that the decision about which muscle type develops centres on a gene-regulatory factor called *Ets1/2* — cells in which the factor is active become heart cells.

But *Ets1/2* is also present, though inactive, in the cells destined to become tail. So something must activate it to set embryonic cells on the path to becoming heart cells. Davidson *et al.* teased apart the genetics to discover that the activating signal comes from a classic growth factor called FGF. Both *Ets1/2* and the FGF signalling pathway have relatives in vertebrates, and these have been



linked to heart development. This implies that the developmental pathway in the sea squirt has been conserved through evolution.

The twist in the tale came from an experiment in which Davidson and colleagues looked at the effect of a permanently active form of *Ets1/2* on embryonic cells destined to become tail. Not only did these cells develop into heart cells, confirming the role of *Ets1/2*, but in some animals the

resulting organ had two chambers rather than one.

So the authors speculate that since it separated from the sea-squirt lineage, an ancestral vertebrate recruited additional heart precursor cells to make a two-chambered heart. As their study shows, even a subtle change in signalling in the pool of cells that can form muscle cells might have allowed this transition.

Helen Dell

K. TEJES/IMAGE QUEST MARINE

The constrictions could propagate from one pericyte to another along the capillary, which could provide a means by which small regions of the vascular system detect and respond to the activity of nearby groups of neurons.

The true impact of pericyte activity in the intact central nervous system still needs to be established. But modifying blood flow through the capillary 'bed' is likely to have a profound impact on overall blood flow. There is *in vivo* evidence from two-photon laser scanning microscopy that capillary blood flow can be dynamically controlled, and even reversed, in some cases^{4,5}. Are these changes at the capillary level due to the actions of pericytes, or do they result from upstream alteration of blood flow in the arterioles? This issue is complicated by the effects of neuronal inputs on blood flow in arterioles and arterioles^{6,7}, or of cells called astrocytes, which can also constrict or dilate vessels⁸⁻¹⁰. Furthermore, localized arteriole constrictions have been observed following transient increases in intracellular calcium in astrocytes, but these were due to contractions of the smooth muscle cells, not to pericyte constriction^{8,9}. A future issue is to find out whether pericytes can also induce constrictions of small arterioles.

Ultimately, the goal is to understand the hierarchy of cerebral blood-flow control by the different elements that can alter vascular dilation and constriction. We will then know which levels are indeed important in inducing the changes in flow that underlie functional

imaging techniques. This knowledge is crucial for understanding the link between regional brain activation and blood flow, and for designing therapies for repairing this relationship in brain diseases such as vascular dementia. The known actors in controlling blood flow are endothelial cells, smooth muscle cells, neuronal inputs and astrocytes. Now, thanks to Peppiatt *et al.*, pericytes can be added to the cast.

Brian A. MacVicar is in the Brain Research Center, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada.

e-mail: bmacvica@interchange.ubc.ca

Michael W. Salter is in the Program in Neurosciences and Mental Health, Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. e-mail: mike.salter@utoronto.ca

- Roy, C. S. & Sherrington, C. S. *J. Physiol. (Lond.)* **11**, 85–108 (1890).
- Lauritzen, M. & Gold, L. *J. Neurosci.* **23**, 3972–3980 (2003).
- Peppiatt, C. M., Howarth, C., Mobbs, P. & Attwell, D. *Nature* **443**, 700–704 (2006).
- Chaigneau, E., Oheim, M., Audinat, E. & Charpak, S. *Proc. Natl Acad. Sci. USA* **100**, 13081–13086 (2003).
- Kleinfeld, D., Mitra, P. P., Helmchen, F. & Denk, W. *Proc. Natl Acad. Sci. USA* **95**, 15741–15746 (1998).
- Edvinsson, L. & Hamel, E. in *Cerebral Blood Flow and Metabolism* (eds Edvinsson, L. & Krause, D. N.) 43–70 (Lippincott Williams & Wilkins, Philadelphia, 2002).
- Cauli, B. *et al. J. Neurosci.* **24**, 8940–8949 (2004).
- Mulligan, S. J. & MacVicar, B. A. *Nature* **431**, 195–199 (2004).
- Metea, M. R. & Newman, E. A. *J. Neurosci.* **26**, 2862–2870 (2006).
- Zonta, M. *et al. Nature Neurosci.* **6**, 43–50 (2003).

is preferentially removed from the system. But apart from a transient spike of ¹³C-enriched rocks following the GOE (the cause of which is still being debated⁷), carbonate rocks that formed before and after 2.4 billion years ago show the same isotopic signature. Why didn't things change when oxygen levels went up?

According to Goldblatt and colleagues⁵, once photosynthesis began, the atmosphere became bistable: it could exist in either a low- or a high-oxygen state. This bistability results from variations in the rate of atmospheric oxygen consumption as oxygen levels change. In a low-oxygen atmosphere, oxygen is rapidly consumed in an ultraviolet-catalysed reaction with biogenic methane. But as oxygen levels increase, so does the concentration of ozone, which shields the atmosphere from solar ultraviolet radiation and thus abates oxygen consumption. Therefore, the atmospheric oxygen budget can change even if other sources and sinks of oxygen remain constant. So maybe the atmosphere got stuck in a low-oxygen state for a long time following the start of photosynthesis, even though oxygen levels were poised to go much higher.

To explain why it is curious that the carbon-isotope signature of rocks did not change after the GOE, one must consider how the long-term oxygen cycle works. Photosynthesis by bacteria (and later by algae and plants) produces oxygen, but it is the burial of these organisms in marine sediments that leaves excess oxygen behind in the atmosphere — this excess oxygen would otherwise be used up as the organisms decay. The organic carbon found in living organisms is depleted in ¹³C relative to ¹²C, so increases in the rate of organic-carbon burial should cause changes in the ratio of carbon isotopes dissolved in the ocean. The isotopic signature of carbonate rocks formed from these dissolved substances should mirror the isotopic changes in sea water.

But carbonate rocks are isotopically similar before and after the GOE (apart from the unexplained spike). In other words, the source of atmospheric oxygen — organic-carbon burial — seems to have remained constant with time, even though atmospheric oxygen levels have changed enormously. This led various authors⁸⁻¹⁰ to propose that the GOE was caused by decreases in the sinks for oxygen — that is, by lower emissions of reduced gases from beneath Earth's crust and by lower discharge rates of dissolved ferrous iron from hydrothermal vents. But all such proposals either have internal inconsistencies or violate other constraints provided by the geological record⁵.

Goldblatt *et al.* now suggest that a mere 3% increase in organic-carbon burial would have been enough to trigger the GOE. Such a change is far too small to be detected in the carbon-isotope record. A decrease in reductant input by the same tiny amount could also have prompted the GOE. Such minor fluctuations could have happened for any number of reasons. However, the authors show that a much larger perturbation is required to cause

EARTH SCIENCES

Ups and downs of ancient oxygen

James F. Kasting

The latest models suggest that atmospheric oxygen could have fluctuated between high and low concentrations once photosynthesis had evolved. But does the geological evidence really support this?

The ancient rise of atmospheric oxygen is of great interest because of its close relationship with evolution, but the geological evidence for this is indirect and subject to interpretation. The consensus for more than 30 years has been that atmospheric oxygen first reached appreciable levels around 2 billion to 2.4 billion years ago^{1,2}, an occasion known as the great oxidation event (GOE). But doubters of this event have remained³.

The GOE story was strengthened considerably by the discovery that minerals in ancient rocks had unusual ratios of sulphur isotopes, a phenomenon known as mass-independent fractionation⁴ (MIF). The only known mechanism that can produce this effect is the break-up of sulphur dioxide by ultraviolet light in a low-oxygen atmosphere. The MIF isotopic signature is small or entirely absent in rocks younger than 2.4 billion years, suggesting that

Earth's atmosphere has been oxygen-rich since that time⁴. In this issue, Goldblatt *et al.* (page 683)⁵ extend the oxygen evolution story, and in doing so may have found some common ground for GOE believers and heretics.

Goldblatt and colleagues⁵ do not challenge the conventional wisdom regarding the rise of oxygen. Instead, they present a model that might resolve two problems that have puzzled geochemists for years. First, why did atmospheric oxygen climb to significant levels only around 2.4 billion years ago, when oxygen-producing bacteria apparently evolved 2.7 billion years ago⁶, or earlier? Second, carbonate rocks that form on the sea floor should acquire a distinctive ratio of carbon isotopes, as organic carbon from photosynthesizing organisms gets buried in marine sediments. Increased organic-carbon burial should cause the ratio of ¹³C to ¹²C in carbonates to rise, because ¹²C

a high-oxygen atmosphere to revert back to a low-oxygen state. This wonderful result might explain why oxygen levels stabilized permanently following the GOE.

But there are other unresolved issues in this saga. There is evidence^{11–13} of low sulphur MIF values in rock formations of between 2.76 billion and 2.92 billion years old, suggestive of high atmospheric-oxygen levels preceding the accepted time of the GOE. So far, no high MIF values have been reported in that time interval. How can this be rationalized?

One possible explanation — the so-called ‘yo-yo’ atmosphere theory — was proposed earlier this year¹¹. This theory suggests that oxygen levels first increased about 3.0 billion years ago, decreasing again about 0.2 billion years later, before their final climb to high concentrations 2.4 billion years ago (the GOE). Goldblatt and colleagues’ model⁵ explicitly predicts that the atmosphere was bistable for some time before the GOE, so maybe the yo-yo theory is correct.

But this would still leave some unexplained observations. For example, the Witwatersrand gold deposits in South Africa contain detrital minerals that were washed down streams between 2.8 billion and 3.0 billion years ago¹⁴. In the presence of oxygen, these minerals should have become oxidized and dissolved. So, either the oxygen levels were never high enough for that, or they repeatedly went up and came back down very quickly. Or perhaps oxygen concentrations did not increase at all, and the low-MIF anomaly seen in post-GOE rocks was produced by some entirely anoxic mechanism, such as the shielding of solar ultraviolet rays by an organic haze^{13,15}.

The jury is still out, but all these contradictory observations are stimulating a lot of creative thinking. Let us hope that this will lead to a more unified understanding of a fascinating era in Earth’s history. The ancient atmosphere may have had a more complex evolution than we imagined.

James F. Kasting is in the Department of

Geosciences, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: kasting@essc.psu.edu

1. Cloud, P. E. *Am. J. Sci.* **272**, 537–548 (1972).
2. Holland, H. D. *The Chemical Evolution of the Atmosphere and Oceans* (Princeton Univ. Press, 1984).
3. Ohmoto, H. *Geology* **24**, 1135–1138 (1996).
4. Farquhar, J. *et al. Science* **289**, 756–758 (2000).
5. Goldblatt, C., Lenton, T. M. & Watson, A. J. *Nature* **443**, 683–686 (2006).
6. Brocks, J. J. *et al. Science* **285**, 1033–1036 (1999).
7. Hayes, J. M. & Waldbauer, J. R. *Phil. Trans. R. Soc. Lond. B* **361**, 931–950 (2006).
8. Kump, L. R., Kasting, J. F. & Barley, M. E. *Geochim. Geophys. Geosyst.* **2**, 2000GC000114 (2001).
9. Catling, D. C. *et al. Science* **293**, 839–843 (2001).
10. Holland, H. D. *Geochim. Cosmochim. Acta* **66**, 3811–3826 (2002).
11. Ohmoto, H. *et al. Nature* **442**, 908–911 (2006).
12. Ono, S. *et al. S. Afr. J. Geol.* (in the press).
13. Peters, M. *et al. Eos Trans. AGU* **86**, 52, Fall Meet. Suppl., Abstr. V21F-06 (2005).
14. Minter, W. E. L. in *Evolution of Early Earth’s Atmosphere, Hydrosphere, and Biosphere — Constraints from Ore Deposits* (Geol. Soc. Am. Mem. Vol. 198) (eds Kesler, S. E. & Ohmoto, H.) 105–109 (2006).
15. Goldman, S. & Kasting, J. F. *Eos Trans. AGU* **86**, 52, Fall Meet. Suppl., Abstr. PP41D-07 (2005).

ECOLOGY

Moving to the ideal free home

Douglas W. Morris

Pike move between two basins of a British lake to maximize their evolutionary fitness. This adaptive behaviour suggests that habitat selection is more significant in population dynamics than was thought.

How should animals choose which habitat to live in? An evolutionary biologist is likely to answer that they should maximize their evolutionary fitness — the likelihood that their genes will be passed on to future generations — and live in the habitat that will produce the most descendants. But there is a catch. If all individuals make the same choice, the habitat will become crowded, the probabilities of survival and reproduction will suffer, and fitness will decline. Other habitats with fewer individuals might be a better option. So animals maximizing fitness through habitat selection will disperse among habitats until no individual can improve its fitness by moving.

This ‘ideal free distribution’¹ assumes that an individual can estimate accurately the fitness that can be attained in different habitats, and is free to move and occupy the ideal place that maximizes fitness. Many ecologists have questioned whether the theory is hamstrung by these preconditions. Can such a model really apply to natural populations? Writing in *Proceedings of the Royal Society*, Thronald Haugen and colleagues² provide convincing evidence that it can. They show that, every year, pike in a northern English lake move from a habitat with low fitness (a higher mortality rate) to one with higher fitness, thus eliminating the initial difference. The research tells us that the

spatial distribution of populations may often represent a dynamic equilibrium caused by habitat selection.

To understand the principle of the ideal free distribution, think of queues at airport security. As a passenger, you want to pass through as quickly as possible. If there are several queues, you choose the one that you think will move

the fastest. But everyone with a boarding pass is trying to do the same. So the number of people in each queue is in dynamic equilibrium because, if one line moves more slowly than others, passengers will swap queues. Thus, at any given time, the average wait before passing through the checkpoint is similar for everyone at the end of the lines.

Now apply the same principle to animals choosing their habitat. At equilibrium, the distribution of individuals among habitats is evolutionarily stable³: no individual can improve its fitness by moving to another habitat. But habitats and population sizes change, disturbing that balance. Individuals will track those changes by moving from one habitat to another until their distribution regains its evolutionary stability. The density in every habitat thus depends on the density in others, just as the length

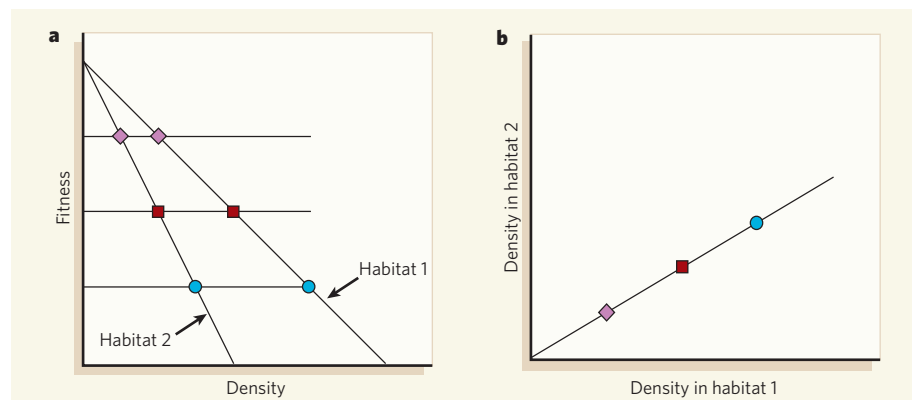


Figure 1 | Testing for an ideal free distribution. **a**, Evolutionary fitness declines more quickly with increasing population density in habitat 2 than it does in habitat 1. Symbols at the intersections with horizontal lines represent densities where fitness is equal in both habitats (an ideal free distribution). **b**, The equilibrium densities can be plotted against each other to yield the expected distribution of individuals in the two habitats (the habitat isodar⁷). Haugen *et al.*² do this for the northern and southern basins of Lake Windermere, and compare these with actual data for pike populations collected since the 1940s. The agreement between the model calculations and the data is astoundingly good.



Figure 2 | Divided lake. Lake Windermere is separated by a shallow sill into two habitats for pike.

of one security queue depends on the others.

There are two ways to test the theory. First, you can do an experiment: you measure the fitness, density and dispersal of a model species living in adjacent habitats. Then you vary the population size; the densities and fitness of individuals in the two habitats should also vary. If, despite changes in density, the mean fitness actually remains equal in each habitat, the animals are following an ideal free distribution.

But what if you can't do the experiment? Then you can stand the theory on its head. Instead of testing whether fitness is equal, you measure fitness at different population sizes in different habitats. You assume that fitness is equalized, and then calculate how many animals should live in each habitat (Fig. 1).

Haugen *et al.*² use an incredible data set to perform this second type of test. For more than 40 years, biologists captured, measured, marked and recaptured pike (*Esox lucius*) in Lake Windermere in northwest England. This lake has two basins (Fig. 2), which are separate pike habitats. Pike reproduce only once each year, so it is appropriate to census their densities and assess habitat selection annually.

The authors use capture histories of individually marked pike to estimate survival and dispersal⁴ of the fish. The probability of observing a given fish depends on the probability of it staying alive between census periods, the probability of it being captured when alive, and the probability of it moving between the two basins. Each of these probabilities depends on such things as the size and sex of the fish, and the basin in which it currently lives. Haugen *et al.* searched for the most likely models⁵ describing the pattern of pike captures in each basin. They merged these models with data on egg production to calculate density-dependent population growth rates in the two habitats. By assuming that these growth rates were equalized through habitat selection, they could back-calculate the number of pike expected in each basin each year. There was a nearly perfect fit between predicted and actual densities.

Thus, pike choosing between the two basins of Lake Windermere behaved as though capable of ideal free habitat selection. Critics

will demand more proof. Do pike disperse to the basin where higher fitness was assured? And, when placed in an experimental environment, do pike adjust their habitat choice to equalize fitness?

Haugen *et al.*² answer both questions. The population density in both basins increases during spawning. Although the southern basin is more productive, pike survival in the south decreases more quickly with population density than it does in the north. Thus, the fitness curves in the two basins diverge (Fig. 1). At an ideal free equilibrium, every fish in each basin will, on average, produce the same number of descendants. But even if the system is in equilibrium after spawning, the different survival rates in each basin cause fitness to rebound more rapidly in the south. The

system will move away from its former equilibrium. So with each passing year — if pike are ideal and free — their net annual dispersal should be biased towards the southern basin.

And it was, most of the time. There was a three-year anomaly when pike reversed their dispersal and headed north. This coincided with a remarkable, serendipitous experiment. Fisheries officials, during only those three years, had reduced the population of pike in the northern basin by approximately 20%.

Most previous tests of ideal free theory have been at a small scale where foragers choose between patches differing in food supply⁶. Pike in Lake Windermere provide compelling evidence that the ideal free distribution occurs at much larger scales where populations are regulated. Dispersal between basins equalizes fitness and homogenizes population growth. Although we need additional tests and experiments at these large scales, the message seems clear: if you study populations, then you must include the potential for habitat selection. ■

Douglas W. Morris is in the Department of Biology, Lakehead University, Thunder Bay, Ontario P7B 5E1, Canada.

e-mail: douglas.morris@lakeheadu.ca

1. Fretwell, S. D. & Lucas, H. L. *Acta Biotheor.* **19**, 16–36 (1969).
2. Haugen, T. O. *et al.* *Proc. R. Soc. Lond. B* doi:10.1098/rspb.2006.3659 (2006).
3. Cressman, R., Křivan, V. & Garay, J. *Am. Nat.* **164**, 473–489 (2004).
4. White, G. C. & Burnham, K. P. *Bird Study* **46**, 120–139 (1999).
5. Burnham, K. P. & Anderson, D. R. *Model Selection and Inference* (Springer, New York, 1998).
6. Tregenza, T. *Adv. Ecol. Res.* **26**, 253–307 (1995).
7. Morris, D. W. *Evol. Ecol.* **2**, 253–269 (1988).

CELL BIOLOGY

Mitochondria shape up

Barbara Conradt

Mitochondria are central to the process of programmed cell death that kills damaged or superfluous cells. Surprisingly, components of the death machinery turn out to be essential for keeping these organelles in shape.

Proteins of the Bcl-2 family are evolutionarily conserved regulators of programmed cell death (apoptosis). However, their mechanisms of action are still not fully resolved^{1,2}. On page 658 of this issue, Youle and colleagues³ report that the mammalian Bcl-2 family members Bax and Bak have additional functions to their apoptotic ones: they control the morphology of mitochondria, the cellular organelles responsible for energy generation. Other components of the apoptotic machinery also perform tasks outside apoptosis^{4,5}. These tasks are unrelated to the proteins' apoptotic roles, however, whereas the non-apoptotic and apoptotic functions of Bax and Bak are linked. This latest study therefore improves our understanding not only of the regulation

of mitochondrial morphogenesis, but also of the apoptotic functions of Bcl-2 proteins.

Bax and Bak are prototypical 'killer' proteins: their pro-apoptotic activities are low in healthy cells but high in cells instructed to die; their overexpression in cultured mammalian cells accelerates apoptosis; and, in mice lacking both Bax and Bak (Bax/Bak DKO mice), most apoptosis is blocked^{1,2}. The ability of Bax and Bak to kill cells is intimately associated with the mitochondria. Whereas these two proteins are found singly in healthy cells, in cells earmarked for apoptosis they tend to group together into oligomers in the outer mitochondrial membrane (OMM). This oligomerization and association with mitochondria has two consequences: it causes the organelles to change

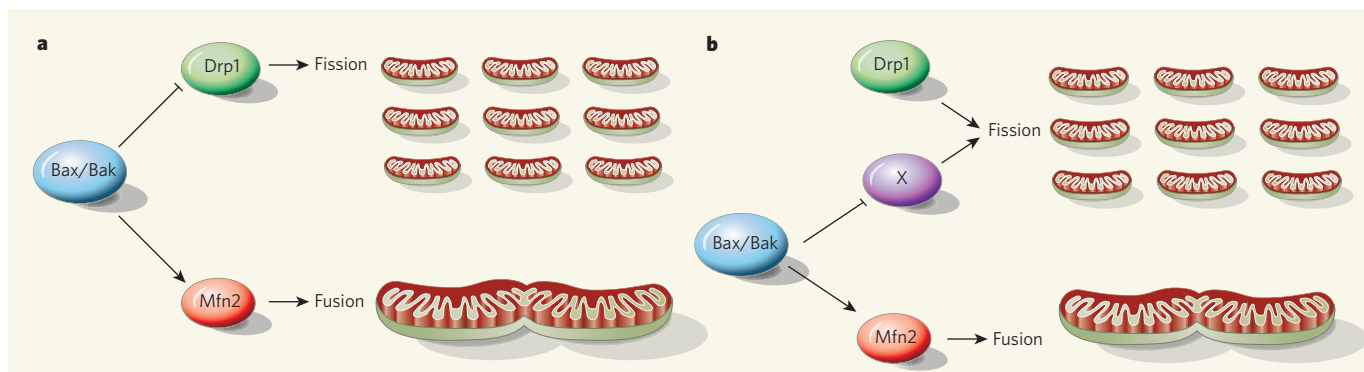


Figure 1 | Regulation of mitochondrial morphogenesis. Youle and colleagues³ show that the proteins Bax and Bak regulate the fusion and fission of mitochondria. Two possible genetic pathways for this regulation are shown. **a**, Bax and Bak promote fusion by activating Mfn2 and inhibiting Drp1. These two proteins are found in the outer mitochondrial membrane together with Bax and Bak, and they are required for mitochondrial fusion and fission, respectively. **b**, Bax and Bak promote fusion by activating Mfn2. An unknown gene, X, acts with Drp1 to cause fission in the absence of Bax and Bak activity.

shape, assuming a fragmented appearance, and it causes the OMM to become permeable to a pool of the cytochrome *c* protein that lies in the mitochondrial intermembrane space^{6–8}. The release of cytochrome *c* from the mitochondria ultimately triggers the start of apoptosis by inducing the assembly of the ‘apoptosome’ — the molecular complex that brings about the cell’s demise.

Mitochondria are highly dynamic organelles that constantly undergo fission and fusion. The Bax- and Bak-dependent fragmentation of the mitochondria in cells about to undergo apoptosis is probably a result of increased fission and decreased fusion^{6,9}. Indeed, in such cells, Bax and Bak gather at foci on the OMM. These sites are probably where future fission will occur, and they also contain the Drp1 and Mfn2 enzymes that are required for mitochondrial fission and fusion, respectively¹⁰. How Bax and Bak control the activities of Drp1 and Mfn2 in cells instructed to die is unknown. Furthermore, how they cause the OMM to become permeable to cytochrome *c* is also a matter of debate. However, one model proposes that channels composed of Bax and Bak proteins form in the OMM^{7,8}.

Youle and colleagues previously found¹¹ that preventing mitochondrial fragmentation in cells destined to die delays the release of cytochrome *c* from the intermembrane space. This suggests that mitochondrial fragmentation promotes the permeabilization of the OMM and/or the release of cytochrome *c*. It is unclear how such a change in mitochondrial morphology accomplishes this, but one theory is that fragmentation aids the formation of Bax and Bak channels in the OMM.

The group³ now shows that Bax and Bak regulate mitochondrial morphogenesis not only in cells instructed to die but also in healthy cells. Unexpectedly, the proteins’ function in this process in healthy cells is opposite to their function in cells about to undergo apoptosis. The authors find that mitochondria in cells from Bax/Bak DKO mice are fragmented as a result of reduced mitochondrial fusion. Therefore, Bax and Bak are required for fusion in healthy cells. Other members of the Bcl-2 family may

also be involved in shaping the mitochondria. Overexpression of two anti-apoptotic cousins of Bax and Bak, Bcl-x_L from mammals and CED-9 from the worm *Caenorhabditis elegans*, promotes mitochondrial fusion in mammalian cells¹². Furthermore, in *C. elegans*, mitochondrial fragmentation can be induced by the fission protein DRP-1, the worm relative of Drp1. This fragmentation is blocked by the loss of function of either the *ced-9* gene or the *egl-1* gene, which encodes a pro-apoptotic cousin of Bax and Bak¹³.

The fragmented mitochondria seen in Bax/Bak DKO cells are suppressed by the overexpression of Mfn2, suggesting that Bax and Bak normally promote fusion by activating Mfn2. Indeed, Youle and colleagues³ find that Bax and Bak are required for various aspects of Mfn2 function, including movement to the foci in the OMM, normal mobility within the OMM, and the ability to form complexes with other proteins. It remains to be determined how Bax and Bak act on Mfn2, but the authors’ results hint that the proteins might interact physically. Interestingly, inactivating the fission protein Drp1 had no effect on fragmented mitochondria in Bax/Bak DKO cells. This indicates that Bax and Bak coordinate both fusion and fission, rather than acting exclusively to promote fusion (Fig. 1).

What Bax and Bak are doing for mitochondria in healthy cells remains to be seen, as, apart from being fragmented, the mitochondria in Bax/Bak DKO cells seem to be fully functional. Another open question is how Bax associates with the OMM in healthy cells, because its association with this membrane in cells instructed to die depends on a structural domain that is normally exposed only in response to apoptotic stimuli. It is feasible that Bax binds to a ‘Bax receptor’ at the OMM, and Mfn2 is one candidate for this receptor. So, too, is Fis1, the established OMM receptor for Drp1; Fis1 affects Bax localization to the OMM in cells earmarked for apoptosis¹⁴.

The newly discovered, non-apoptotic Bax function depends on a domain of the Bax protein, the BH3 domain, that is required for apoptosis¹². Hence, the mechanisms by which

Bax and Bak regulate mitochondrial morphogenesis in healthy cells and in cells instructed to die might be related. However, as mentioned above, the function of Bax and Bak in regulating mitochondrial morphogenesis in healthy cells is the opposite of their function in cells about to undergo apoptosis. It will thus be necessary to discover whether and how apoptotic stimuli induce this functional switch in Bax and Bak. Possible culprits for the switching mechanism are the ‘BH3-only proteins’¹⁵, which can bind to Bax and Bak in cells instructed to die, inducing a conformational change in Bax and Bak. The levels of Bax and Bak associated with the OMM might also be significant, as the overexpression of Bax in healthy cells can induce mitochondrial fragmentation and be sufficient to trigger the switch between the ‘healthy’ and ‘death-dealing’ Bax functions^{11,16}. Whatever the answers to these questions, one thing is certain: the relationship between mitochondria and the apoptotic machinery is more complex than previously thought.

Barbara Conradt is in the Department of Genetics, Dartmouth Medical School, 7400 Remsen, Hanover, New Hampshire 03755, USA. e-mail: barbara.conradt@dartmouth.edu

- Danial, N. N. & Korsmeyer, S. J. *Cell* **116**, 205–219 (2004).
- Cory, S. & Adams, J. M. *Nature Rev. Cancer* **2**, 647–656 (2002).
- Karbowksi, M., Norris, K. L., Cleland, M. M., Jeong, S.-Y. & Youle, R. J. *Nature* **443**, 658–662 (2006).
- Woo, M. et al. *Nature Immunol.* **4**, 1016–1022 (2003).
- Huang, D. C., O’Reilly, L. A., Strasser, A. & Cory, S. *EMBO J.* **16**, 4628–4638 (1997).
- Youle, R. J. & Karbowksi, M. *Nature Rev. Mol. Cell Biol.* **6**, 657–663 (2005).
- Wang, X. *Genes Dev.* **15**, 2922–2933 (2001).
- Newmeyer, D. D. & Ferguson-Miller, S. *Cell* **112**, 481–490 (2003).
- Okamoto, K. & Shaw, J. M. *Annu. Rev. Genet.* **39**, 503–536 (2005).
- Karbowksi, M. et al. *J. Cell Biol.* **159**, 931–938 (2002).
- Frank, S. et al. *Dev. Cell* **1**, 515–525 (2001).
- Delivani, P., Adrain, C., Taylor, R. C., Duriez, P. J. & Martin, S. J. *Mol. Cell* **21**, 761–773 (2006).
- Jagasia, R., Grote, P., Westermann, B. & Conradt, B. *Nature* **433**, 754–760 (2005).
- Lee, Y. J., Jeong, S. Y., Karbowksi, M., Smith, C. L. & Youle, R. J. *Mol. Biol. Cell* **15**, 5001–5011 (2004).
- Fletcher, J. I. & Huang, D. C. *Cell Death Differ.* **13**, 1268–1271 (2006).
- Desagher, S. & Martinou, J. C. *Trends Cell Biol.* **10**, 369–377 (2000).

BRIEF COMMUNICATIONS

An accessory chromophore in red vision

The rods in salamanders' retinas can co-opt a molecule derived from chlorophyll to detect red light.

In the absence of a red-sensitive visual pigment, some deep-sea fish use a chlorophyll derivative in their green-sensitive rod cells in order to see deep-red light^{1–3}. Here we show that living rods extracted from a salamander can also accumulate an exogenous chlorophyll derivative, chlorin *e*₆, that renders them as sensitive to red light as they are to green. This vision enhancement by an unbleachable chlorophyll derivative might therefore be a general phenomenon in vertebrate photoreception.

The outer segments of rod cells are rich in rhodopsin: this visual pigment consists of retinal, a form of vitamin A that isomerizes when it absorbs light, and opsin, a protein that changes shape in response to retinal isomerization, which triggers a signal.

We measured the absorbance spectra of individual rods from tiger salamanders (*Ambystoma tigrinum*) by microspectrophotometry (for methods, see supplementary information). Absorbance peaked at wavelengths of 520 nm and 278 nm, owing to the retinal chromophore and the opsin, respectively.

When rods were preincubated with chlorin (Fig. 1a), we found that chlorin targeted the rhodopsin-laden outer segments of the rods; it increased absorbance at 278 nm, and added peaks at 402 nm and 668 nm (Fig. 1b). The peak at 520 nm was unchanged, so if chlorin did bind to rhodopsin, it did so without influencing the retinal chromophore or the binding pocket that spectrally tunes rhodopsin. The slight increase in absorbance of 278-nm light indicates that chlorin may be perturbing aromatic residues on the surface of rhodopsin.

We selected cells with the highest chlorin content (chlorin:rhodopsin ratio was 0.77 ± 0.22 , mean $\pm$ s.e.m.; $n = 3$) to examine chlorin's effect on rhodopsin's sensitivity to light, measured by its bleaching rate. The bleaching rate was unchanged at 560 nm, a wavelength absorbed well by rhodopsin but poorly by chlorin. However, it was accelerated 180-fold at 668 nm, where chlorin absorbs strongly but rhodopsin absorbance is negligible (Fig. 1c).

This bleaching enhancement was tens to

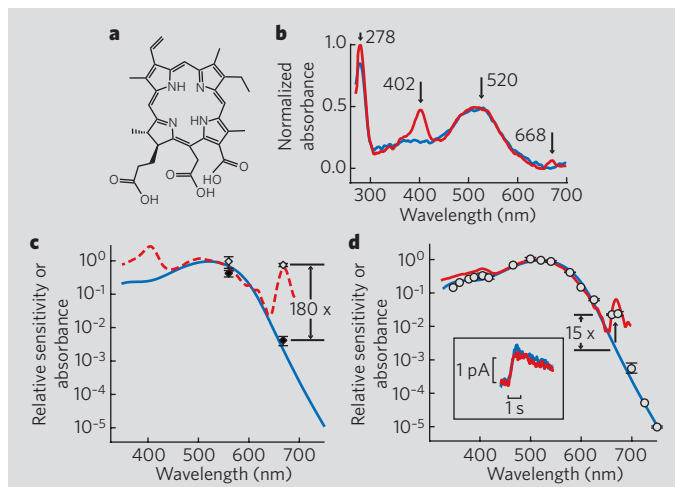


Figure 1 | Chlorin-enhanced response to red light. **a**, Structure of the tetrapyrrole chlorin *e*₆. **b**, Averaged absorption from chlorin-treated (red trace; $n = 5$) and control (blue trace; $n = 5$) rods. The spectrum of treated rods is normalized to the 278-nm peak, and for untreated rods is scaled to match the 520-nm peak for treated rods. **c**, Rhodopsin bleaching from untreated rods (filled symbols; mean $\pm$ s.e.m.), aligned to a template fit⁷ to the spectral sensitivity of seven untreated rods (blue trace). The bleaching rate of rods with high chlorin content (open symbols) is enhanced at 668 nm but not at 560 nm; the rate mirrors the average absorption of such rods (dashed trace), showing that quanta absorbed by chlorin and by rhodopsin are equally effective at bleaching rhodopsin (for details, see supplementary information). **d**, Spectral sensitivity (circles; $n = 3$) increases at 661 and 672 nm in chlorin-treated rods and matches the normalized absorption of a set of rods treated with chlorin and measured microspectrophotometrically (red trace); blue trace is from **c**. Inset, response (in picoamperes, pA) of a chlorin-treated rod to dim flashes at 540 nm (blue; 6.4×10^{-1} photons μm^{-2}) and at 672 nm (red; 22.3 photons μm^{-2}).

hundreds of times greater than the effect of chlorophyll derivatives on dragon fish (*Malacosteus niger*)^{2,3} and bovine⁴ rhodopsins. This shows that chlorophyll-sensitized vision is feasible. Adding 9-*cis*-retinal regenerated a light-sensitive pigment, irrespective of the bleaching conditions that had been used, indicating that chlorin-mediated rhodopsin bleaching was not due to a permanent, photo-induced degradation of opsin (for details, see supplementary information).

In electrical recordings of rods exposed to chlorin, responses to dim flashes were similar in form at all wavelengths (Fig. 1d; inset shows response to two wavelengths). This indicates that photons absorbed by chlorin or by the retinal chromophore directly induce the same active state in rhodopsin. Chlorin treatment increased rod sensitivity to red light (at 661 and 672 nm) by 15-fold (Fig. 1d). This enhancement was less than the 180-fold increase in bleach-

ing rate found earlier, because the cells used for recordings were randomly chosen and contained about 20 times less chlorin (chlorin:rhodopsin ratio, 0.05 ± 0.01 , mode of 61 rods $\pm$ s.e.m.; $n = 15$). Nevertheless, in both microspectrophotometric and electrophysiological experiments, chlorin-assisted bleaching was almost as efficient as direct photoisomerization of the native chromophore (Fig. 1, and see supplementary information).

The mechanism behind the highly efficient energy transfer from chlorin to the retinal chromophore is unclear at this point. However, our results on an amphibian, together with findings from dragon fish^{1–3}, bovine⁴, insect⁵ and prokaryotic⁶ rhodopsins, highlight a potential for spectral enhancement by accessory chromophores in all opsin-based photoreception systems.

T. Isayama*, **D. Alexeev***,
C. L. Makino*, **I. Washington†**,
K. Nakanishi†, **N. J. Turro†**

*Department of Ophthalmology, Massachusetts Eye & Ear Infirmary, Boston, Massachusetts 02114, USA
e-mail: cmakino@meei.harvard.edu

†Department of Chemistry, Columbia University, New York, New York 10027, USA

1. Bowmaker, J. K., Dartnall, H. J. A. & Herring, P. J. *J. Comp. Physiol. A* **163**, 685–698 (1988).
2. Douglas, R. H. *et al. Nature* **393**, 423–424 (1998).
3. Douglas, R. H. *et al. Vision Res.* **39**, 2817–2832 (1999).
4. Washington, I., Brooks, C., Turro, N. J. & Nakanishi, K. *J. Am. Chem. Soc.* **126**, 9892–9893 (2004).
5. Kirschfeld, K. & Vogt, K. *Vision Res.* **26**, 1771–1777 (1986).
6. Balashov, S. P. *et al. Science* **309**, 2061–2064 (2005).
7. Govardovskii, V. I., Fyhrquist, N., Reuter, T., Kuzmin, D. G. & Donner, K. *Visual Neurosci.* **17**, 509–528 (2000).

Supplementary information accompanies this communication on Nature's website.

Received 4 April; accepted 25 August 2006.

Competing financial interests: declared none.
doi:10.1038/443649a

Corrigendum

Developmental technology: Dogs cloned from adult somatic cells
Byeong Chun Lee *et al. Nature* **436**, 641 (2005)

The name of the seventh author on this communication was misspelled. His correct name is M. Shamim Hossein.

Phosphoinositides in cell regulation and membrane dynamics

Gilbert Di Paolo¹ & Pietro De Camilli²

Inositol phospholipids have long been known to have an important regulatory role in cell physiology. The repertoire of cellular processes known to be directly or indirectly controlled by this class of lipids has now dramatically expanded. Through interactions mediated by their headgroups, which can be reversibly phosphorylated to generate seven species, phosphoinositides play a fundamental part in controlling membrane-cytosol interfaces. These lipids mediate acute responses, but also act as constitutive signals that help define organelle identity. Their functions, besides classical signal transduction at the cell surface, include regulation of membrane traffic, the cytoskeleton, nuclear events and the permeability and transport functions of membranes.

In the early 1950s, Hokin and Hokin first observed changes in the turnover of membrane phospholipids on stimulation of exocrine tissues. This so-called 'phospholipid effect' was subsequently detected in a variety of tissues and attributed to changes in the turnover of phosphatidylinositol (PtdIns) and its phosphorylated products, called phosphoinositides¹. Studies in this area eventually focused on the function of phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P₂) in signal transduction as a precursor of intracellular messengers generated by phospholipases². Unexpected twists came from the discovery of direct signalling roles for PtdIns(4,5)P₂ (refs 3–5) and of phosphoinositides phosphorylated at the 3 position^{6–8}. Furthermore, genetic screens in yeast and studies in evolutionarily higher eukaryotic cells demonstrated the involvement of phosphoinositide-metabolizing enzymes in membrane traffic as well as cytoskeleton organization^{9–12}. The discovery and functional characterization of other rare phosphoinositides^{13,14}, phosphoinositide-binding protein modules^{15,16} and a plethora of phosphoinositide-metabolizing enzymes^{17–19} have now dramatically advanced this field. Phosphoinositides have been implicated in nearly all aspects of cell physiology.

The importance of the signalling and constitutive functions of phosphoinositides is paralleled by that of soluble inositol polyphosphates (InsPs). In addition to the fundamental role of Ins(1,4,5)P₃ in intracellular Ca²⁺ signalling², many functions have been described for other InsPs, including gene transcription, RNA editing, nuclear export and protein phosphorylation. Although InsPs and phosphoinositides are metabolically and functionally interconnected, this review will focus selectively on the role of phosphoinositides.

Metabolism and spatial distribution of phosphoinositides

Inositol phospholipids are concentrated at the cytosolic surface of membranes. PtdIns, the precursor of phosphoinositides, is synthesized primarily in the endoplasmic reticulum (ER) and is then delivered to other membranes either by vesicular transport or via cytosolic PtdIns transfer proteins. Reversible phosphorylation of its inositol ring at positions 3, 4 and 5 results in the generation of seven phosphoinositide species (Fig. 1a). PtdIns typically represents less than 15% of the total phospholipids found in eukaryotic cells and phosphoinositides are generally less abundant by one order of magnitude, with PtdIns(4)P and PtdIns(4,5)P₂ representing the

bulk of these lipids in mammalian cells. Each of the seven phosphoinositides has a unique subcellular distribution with a predominant localization in subsets of membranes (Fig. 1b, c; see below). Furthermore, within a given membrane, the localization of specific phosphoinositides can be heterogeneous. Pools of phosphoinositides have also been reported in the nucleus²⁰. The differential intracellular distribution of phosphoinositides, together with their high turnover, makes these lipids optimal mediators of signalling events in all cellular compartments.

Control of the membrane-cytosol interface

Phosphoinositides achieve direct signalling effects through the binding of their head groups to cytosolic proteins or cytosolic domains of membrane proteins. Thus, they can regulate the function of integral membrane proteins, or recruit to the membrane cytoskeletal and signalling components. This action is reminiscent of that of phosphotyrosine residues of membrane proteins, which mediate the recruitment of phosphotyrosine-binding modules (for example, SH2 domains). Typically, binding of proteins to phosphoinositides involves electrostatic interactions with the negative charges of the phosphate(s) on the inositol ring. In some cases, adjacent hydrophobic amino acids strengthen the interaction through a partial penetration into the bilayer^{15,16}. Protein surfaces that interact with phosphoinositides can consist either of clusters of basic residues within unstructured regions, such as those found in many actin regulatory proteins (for example, profilin)⁵, or of folded modules, such as the pleckstrin homology domain (Fig. 1d and see below)^{15,16,21}. The distinction between these two types of interaction can be blurred. Furthermore, unstructured peptides can undergo folding on binding to phosphoinositide head groups. The number of phosphoinositide-binding modules is rapidly expanding (Fig. 1d) and their differential properties greatly amplify the signalling potential of phosphoinositides^{15,16,21}.

Coincidence detection code and organelle identity

Generally, the interaction between phosphoinositides and cytosolic proteins is of low affinity. Higher affinity, and thus more stable, membrane-protein interactions are produced when phosphoinositides cooperate with one or more additional binding sites within

¹Department of Pathology and Cell Biology, The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, New York, New York 10032, USA. ²Departments of Cell Biology and Neurobiology, Program in Cellular Neuroscience, Neurodegeneration and Repair, Howard Hughes Medical Institute and Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, Connecticut 06510, USA.

the membrane. Owing to the heterogeneous subcellular localization of phosphoinositides, this dual (or multiple) key-based recognition mechanism generates a powerful coincidence detection code for the regulation of membrane–cytosol interactions and thus for the definition of organelle identity (Fig. 2a)^{22,23}. For example, although the clathrin adaptors AP-1 and AP-2 bind some of the same cargo proteins, they are recruited to the Golgi complex or the plasma membrane,

respectively, at least in part owing to the different phosphoinositide content of these membranes (Fig. 2b)^{24–27}. One advantage of this dual code is that each component of the code can be independently regulated (for example, by lipid enzyme activation or co-receptor phosphorylation), thus offering an opportunity for signal integration.

The spatial restriction and steady-state levels of specific phosphoinositides are determined primarily by the concerted action of phosphoinositide kinases and phosphatases, whose localization is tightly controlled. Phosphoinositide segregation on different membranes helps ensure vectorial membrane traffic, which is critical to achieve transport of cargo from one compartment to another. For example, a carrier vesicle must harbour a lipid composition that promotes the recruitment to its surface of factors needed for fusion with its acceptor membrane (Fig. 2c). This state must be maintained until fusion has occurred and terminated thereafter. Conversely, the acceptor membrane must have properties that make it suitable as a partner for fusion and as a template for the recruitment of budding factors needed for the next trafficking step. After budding, these factors (generally coat proteins) must be shed. Modifications in phosphoinositide environment after fusion (by a dilution effect and exposure to a different enzymatic context) and then after budding (for example, by the selective recruitment of a phosphoinositide phosphatase onto the vesicle) can efficiently account for these global changes because a new phosphoinositide would generate a new identity code (Fig. 2c).

A phosphoinositide-based code represents a very effective way to define organelle identity. Phospholipids can rapidly diffuse within, but not between, membranes (unless assisted by transfer proteins). In addition, phosphoinositides can be rapidly interconverted from one species to another by strategically localized kinases and phosphatases as a membrane carrier translocates from one compartment to the next (Figs 1c and 2c). Several enzymes contain modules that bind the substrate phosphoinositide (Fig. 1d), thus facilitating recruitment of enzymes to the appropriate membrane and their retention at this membrane for catalysis as long as substrates to be consumed are present.

Cooperation with small GTPases

Another class of molecules that mediate or enable the recruitment of cytosolic proteins to specific membrane compartments or subcompartments is the Ras superfamily of small GTPases. These include proteins implicated in signalling (for example, the Ras family itself), actin regulation (for example, the Rho family), organelle and vesicular transport (for example, the Arf, Arl and Rab families)²³. Small GTPases shuttle between a GDP (inactive)- and a GTP

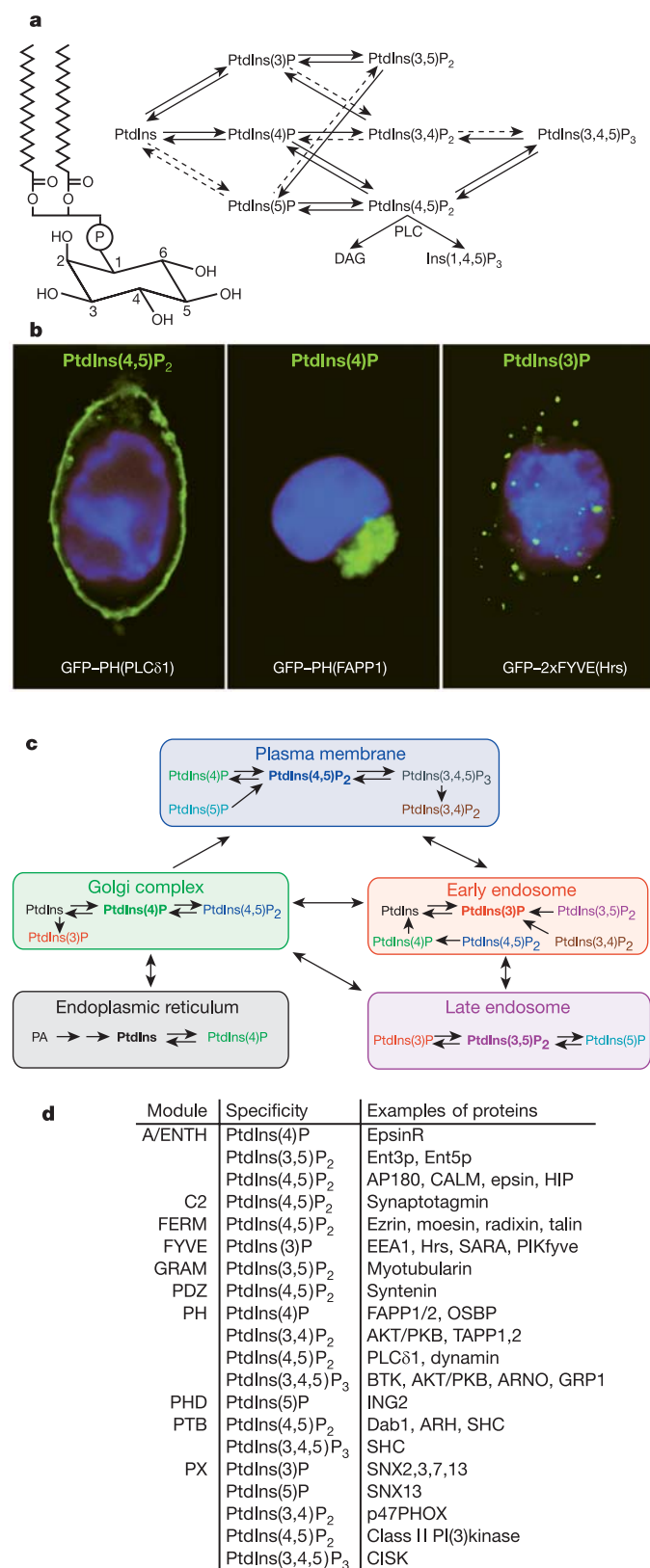


Figure 1 | Metabolism and subcellular distribution of phosphoinositides.

a, Metabolic reactions leading to the generation of seven phosphoinositide species from PtdIns. Reactions indicated with dotted arrows have been shown *in vitro*, but their importance in living cells remains unclear. DAG, diacylglycerol. **b**, Fluorescence micrographs illustrating the predominant localization of PtdIns(4,5)P₂ (plasma membrane), PtdIns(4)P (Golgi) and PtdIns(3)P (endosomes) in Chinese hamster ovary cells as revealed by the pleckstrin homology (PH) domain of PLCδ1, the PH domain of FAPP1 and the tandem FYVE domain of Hrs (all fused to GFP, green). The nucleus is shown in blue (DAPI staining). **c**, Predominant subcellular localization of phosphoinositide species. PtdIns(4,5)P₂ and PtdIns(3,4,5)P₃ are concentrated at the plasma membrane, possibly enriched in raft-like structures. PtdIns(3,4)P₂ is mostly found at the plasma membrane and in the early endocytic pathway. PtdIns(4)P is enriched at the Golgi complex, but also present at the plasma membrane. PtdIns(3)P is concentrated in early endosomes and PtdIns(3,5)P₂ on late compartments of the endosomal pathway. The location of PtdIns(5)P remains poorly characterized. Phosphorylation/dephosphorylation events occurring on each subcellular compartment are illustrated. In the ER, the conversion from phosphatidic acid (PA) to PtdIns involves more than one reaction. **d**, Phosphoinositide-binding modules, their binding preference and some examples of proteins that contain them.

(active)-bound state. When bound to GTP, they interact with a variety of effectors. This cycle is regulated by guanine nucleotide exchange factors (GEFs) and GTPase activating proteins (GAPs), which stimulate GTP loading and hydrolysis, respectively. Not surprisingly, many examples of a close functional relationship between phosphoinositide metabolism and these proteins have been reported (Fig. 3). First, phosphoinositides regulate the recruitment of GAPs and GEFs to membranes and their activity, given the presence of phosphoinositide-binding modules in many of these factors (Fig. 3a). Second, several phosphoinositide-metabolizing enzymes are effectors

of GTPases, thus providing the basis for positive and negative feedback loops^{22,28,29} (Fig. 3b). Third, membrane-bound GTPases often function as co-receptors with phosphoinositides in the recruitment of cytosolic proteins to specific subcompartments (Fig. 3c). Because mammalian cells express more than 100 small GTPases, the latter mechanism dramatically increases (in a combinatorial manner) the repertoire of identity codes for membranes and membrane domains^{22,23}.

The following sections provide an overview of some of the most important regulatory roles of phosphoinositides in the major cell compartments.

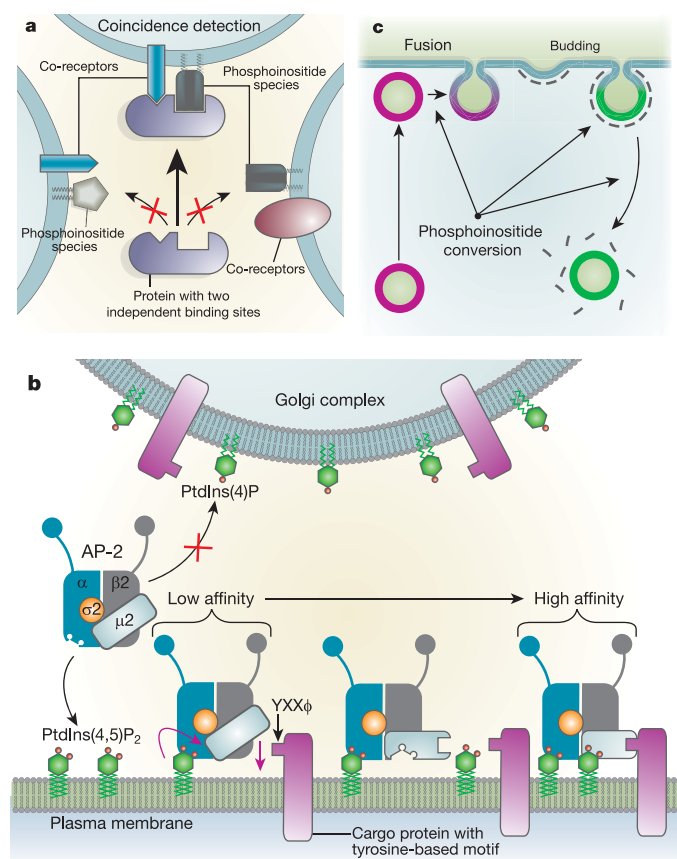
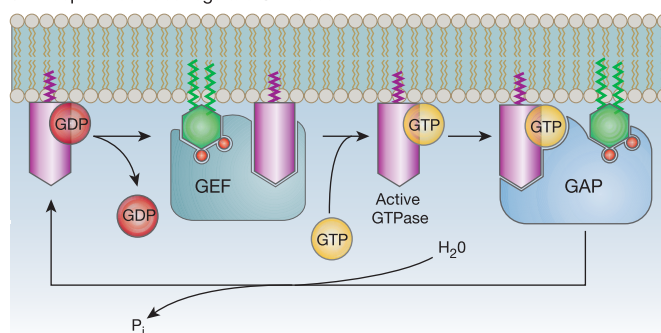


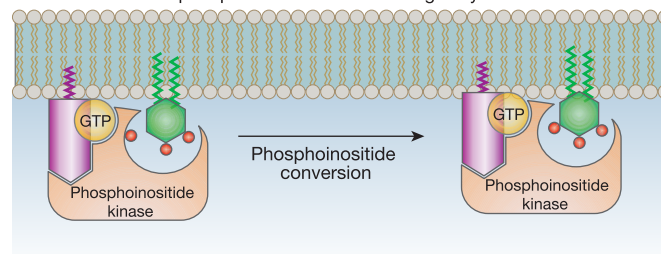
Figure 2 | Coincidence detection in phosphoinositide signalling.

a, Phosphoinositides mediate the recruitment of proteins to membranes using their phosphorylated head groups. A high specificity is achieved when other signals act in concert with phosphoinositides. This dual-key mechanism plays an important part in the spatial and temporal regulation of the interactions between cytosolic proteins and membranes. **b**, To illustrate the concept of coincidence detection, the case of the clathrin adaptor AP-2 is shown. This protein complex, which consists of four subunits (α , β 2, μ 2 and σ 2), is selectively recruited to the plasma membrane owing, in part, to its ability to bind PtdIns(4,5)P₂. It has been proposed that an initial contact with this membrane mediated by the PtdIns(4,5)P₂ binding site present on the α subunit triggers a conformational change in the μ 2 subunit. This change, in turn, leads to the exposure of an additional binding site for PtdIns(4,5)P₂ and of a binding site for membrane protein co-receptors, such as cargo proteins harbouring tyrosine-based motifs (YXX ϕ , where ϕ denotes bulky, hydrophobic residues). These multiple interactions, which can also be modulated by the phosphorylation of AP-2, cooperate in stabilizing the association of AP-2 with the plasma membrane. Although cargo proteins are also present in intracellular membranes, AP-2 is not efficiently recruited to such membranes owing to the lack of a cognate phosphoinositide²⁶. **c**, A schematic drawing illustrating changes in phosphoinositide composition (represented by changes in membrane colour) that accompany fusion and budding reactions. On fusion, a vesicle will acquire the predominant phosphoinositide composition of the target membrane. Conversely, membrane budding generates a new closed compartment whose phosphoinositide composition can be rapidly and globally changed.

a Phosphoinositides regulate GTPases



b GTPases control phosphoinositide-metabolizing enzymes



c Phosphoinositides and GTPases act as co-receptors

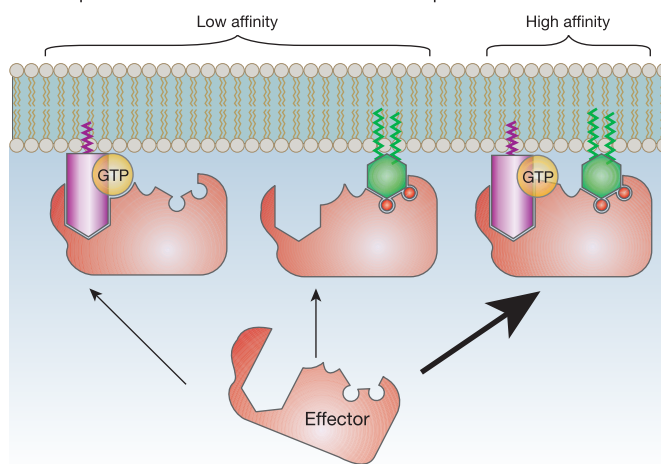


Figure 3 | Cartoons illustrating examples of functional interplay between phosphoinositides and small GTPases. **a**, GTPases (purple) shuttle between GDP (inactive)- and GTP (active)-bound states, the interconversion of which is regulated by GEFs and GAPs. Phosphoinositides (green) regulate the nucleotide cycle of GTPases owing to their ability to bind and activate both GEFs and GAPs. **b**, Many small GTPases bind (and often regulate) phosphoinositide-metabolizing enzymes (here, a phosphoinositide kinase is depicted), thereby contributing to the generation of membrane domains enriched in specific phosphoinositides. **c**, Phosphoinositides and small GTPases act in concert to recruit effectors to specific membrane compartments. These three mechanisms (**a**, **b**, **c**) can occur simultaneously on the same membrane compartment, thereby giving rise to complex feedback loops.

Classical signal transduction at the plasma membrane

PtdIns(4,5)P₂, which is concentrated in the plasma membrane, participates in nearly all events that occur at, or involve, the cell surface (Fig. 4). In addition, it plays a major part in the transduction of extracellular signals, either via its metabolites, or through fluctuations of its own levels.

PtdIns(4,5)P₂ is produced by the phosphorylation of either PtdIns(4)P (by type I PtdInsP kinases) or PtdIns(5)P (by type II PtdInsP kinases)^{5,14,19}. There is only very little PtdIns(5)P in cells and most cell surface PtdIns(4,5)P₂ is generated from PtdIns(4)P. This lipid is delivered to the cell surface by membrane carriers derived from the Golgi complex and from recycling organelles (see below), but is also produced locally at the plasma membrane¹⁸. PtdIns(4,5)P₂ hydrolysis is due to both phospholipases and phosphatases with fundamentally different physiological consequences. Cleavage by phospholipases, such as phospholipase C (PLC) and phospholipase A₂ (PLA₂), gives rise to metabolites that propagate and amplify signalling, whereas dephosphorylation, primarily by 5-phosphatases, controls steady-state levels of PtdIns(4,5)P₂ and turns off its signalling. For example, during endocytosis, dephosphorylation removes PtdIns(4,5)P₂ from internalized membranes, thus restricting the localization of PtdIns(4,5)P₂ to the cell surface without coupling the endocytic reaction to the generation of signalling metabolites³⁰.

Another fate of PtdIns(4,5)P₂ is conversion to PtdIns(3,4,5)P₃ primarily by class I phosphoinositide (3)kinases (PI(3) kinases)^{31–33}. This lipid is present in negligible amounts in resting cells, but can transiently and dramatically increase in response to growth factor stimulation. Its regulatory role is of fundamental importance in the physiology of higher eukaryotes, as it mediates a large variety of effects, including cell proliferation, migration, chemotaxis, phagocytosis and macro-pinocytosis, differentiation, survival and metabolic changes^{31–33}. PtdIns(3,4,5)P₃ acts by recruiting effectors that activate synergistic signalling pathways³¹. Prominent among the PtdIns(3,4,5)P₃ target proteins are GEFs and GAPs for small GTPases that function as regulators of the actin cytoskeleton, and the protein kinases PDK and AKT/PKB, which cooperate in the activation of important signalling cascades, such as the Tor pathway^{31,32}. PtdIns(3,4,5)P₃ is dephosphorylated by two types of enzymes, with different outcomes. Dephosphorylation at the 3 position by PTEN is the 'off' signal³⁴. In contrast, dephosphorylation by 5-phosphatases, such as SHIP-1 and 2, generates PtdIns(3,4)P₂, which shares several actions with PtdIns(3,4,5)P₃ (in addition to unique signalling properties) and may prolong the duration of PtdIns(3,4,5)P₃ signalling¹⁷. This has major relevance to medicine, given the importance of PtdIns(3,4,5)P₃ metabolism imbalance in diabetes³² and cancer³¹. PTEN is also a potent tumour suppressor (Table 1)^{34,35}.

Regulation of integral plasma membrane proteins

An emerging theme in phosphoinositide signalling is the involvement of PtdIns(4,5)P₂ in the regulation of integral membrane proteins at the plasma membrane. These include a variety of inward rectifier and voltage-gated potassium channels, calcium channels and pumps, transient receptor potential channels, epithelial sodium channels, ion exchangers^{36,37} and probably, many other classes of proteins, such as receptors for chemical ligands. For some proteins, presence of binding sites for PtdIns(4,5)P₂ may represent a

mechanism for optimal function in their appropriate cellular context, that is, the plasma membrane. In other cases, regulation by PtdIns(4,5)P₂ may allow these proteins to function as effectors of receptors coupled to PtdIns(4,5)P₂ synthesis or degradation. Regulatory actions of PtdIns(4,5)P₂ are well documented for several transient receptor potential channels, many of which function in sensory transduction^{36,37}.

Plasma membrane–cytoskeleton interactions

PtdIns(4,5)P₂ and PtdIns(3,4,5)P₃, often in concert with small GTPases, participate in the recruitment and activation of a wide variety of actin regulatory proteins at the plasma membrane, thereby controlling cell shape, motility, cytokinesis and a wide variety of other processes. A well-characterized mechanism governing actin polymerization and involving PtdIns(4,5)P₂ is the ARP2/3-mediated nucleation of actin networks. For example, PtdIns(4,5)P₂, in cooperation with the small GTPase Cdc42, binds N-WASP and triggers a conformational change that allows its binding to and activation of the ARP2/3 complex^{38,39}. Other activators of N-WASP (for example, Toca) are also recruited to the plasma membrane by the cooperative action of phosphoinositides and Cdc42, thus providing mechanisms for the amplification of PtdIns(4,5)P₂ action⁴⁰.

PtdIns(4,5)P₂ facilitates actin filament elongation at the plus ends (that is, the actin filament tip at which addition of G-actin monomers preferentially occurs) by promoting the dissociation of capping proteins, such as gelsolin and CapZ³. In addition, through its binding to profilin, it promotes the dissociation of actin-monomer–profilin complexes, thus making G-actin available^{5,38}.

Binding to PtdIns(4,5)P₂ is also critical for the function of proteins that act as adaptors between the membrane and the actin cytoskeleton at sites of cell–matrix or cell–cell adhesion. These adaptors include the band 4.1/ezrin/radixin/moesin (FERM) domain-containing proteins. Typically, PtdIns(4,5)P₂ binds to the FERM domain of these proteins, thus releasing them from an autoinhibited state. The FERM domain of talin not only binds PtdIns(4,5)P₂, which enhances its affinity for integrins, but also the PtdIns(4,5)P₂-synthesizing enzyme PtdInsP kinase type 1 γ and so helps generate this lipid at sites of cell adhesion^{41,42}.

PtdIns(3,4,5)P₃ mediates the effect of growth factors on the formation of the peripheral ruffles implicated in cell migration^{32,43}. Mechanistically, the actions of PtdIns(3,4,5)P₃ are similar to those of PtdIns(4,5)P₂, although they often involve a distinct set of effector proteins. For example, WAVE, a WASP-related protein, is specifically recruited and activated by PtdIns(3,4,5)P₃ (ref. 43). In *Dictyostelium discoideum*, the spatial restriction of PtdIns(3,4,5)P₃ at the leading edge is responsible for polarized cell migration during chemotaxis⁴⁴.

The regulatory function of phosphoinositides on the cytoskeleton is not restricted to actin. For instance, PtdIns(4,5)P₂-enriched microdomains in the plasma membrane participate in the regulation of microtubule plus-end capture and stabilization, a phenomenon required for normal polarized motility⁴⁵. In another example, a microtubular motor was shown to be anchored to cell organelles using a phosphoinositide-binding pleckstrin homology domain, thus implicating phosphoinositides in microtubule-based intracellular motility⁴⁶.

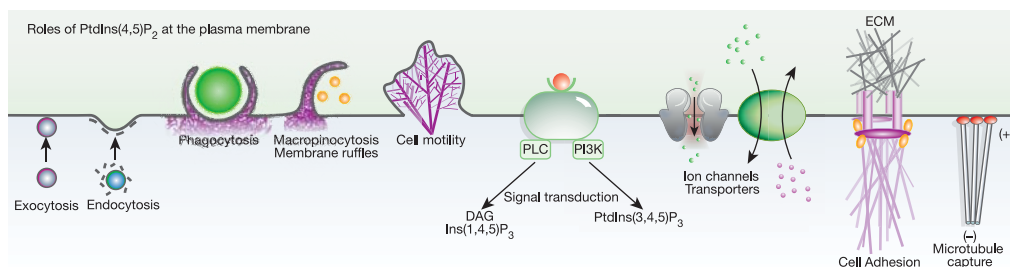


Figure 4 | Examples of processes regulated by PtdIns(4,5)P₂ at the plasma membrane. See text for explanations. ECM, extracellular matrix; PI3K, PI(3)kinase.

Exocytosis

Plasma membrane PtdIns(4,5)P₂ is directly implicated in exocytosis owing to a critical role at a pre-fusion stage (that is, ‘vesicle priming’). It acts both in *cis* on plasma membrane proteins (such as CAPS, a priming factor in neuroendocrine cells) and in *trans* on vesicle proteins (such as synaptotagmin, a Ca²⁺ sensor for exocytosis)^{12,47}. Importantly, in-*trans* interactions probably cooperate with SNARE pairing⁴⁸ in marking the plasma membrane as the appropriate partner for the fusion of these organelles. Generally, levels of PtdIns(4,5)P₂ in the plasma membrane positively correlate with the size of the readily releasable pool of secretory vesicles (that is the pool of membrane-docked vesicles that is competent for exocytosis)^{49–51}. PtdIns(4,5)P₂ also indirectly affects exocytosis through metabolites produced by PLC-dependent cleavage: diacylglycerol is a critical ligand for the priming factor Munc13 (refs 52, 53), whereas Ins(1,4,5)P₃ mediates calcium responses that modulate and in some cases trigger secretion.

Endocytosis

PtdIns(4,5)P₂ has an important role in all forms of endocytosis, because it functions as an important co-receptor for the recruitment and regulation of endocytic proteins selectively to the plasma membrane^{22,27}. PtdIns(4,5)P₂ (in most cases also PtdIns(3,4,5)P₃) binds all the known endocytic clathrin adaptors (for example, AP-2, AP180/CALM, epsin) and many other endocytic factors, including dynamin, which controls the fission reaction^{22,24,27}. Other actions of PtdIns(4,5)P₂ in endocytosis reflect its effects on the actin cytoskeleton, which is implicated in all internalization pathways⁵⁴.

Specific retention of PtdIns(4,5)P₂ and PtdIns(3,4,5)P₃ at the plasma membrane and rapid dissociation of endocytic factors following internalization require the removal of these phosphoinositides from endocytic membranes. This function is largely fulfilled by inositol(5)phosphatases, as highlighted by studies on synaptojanin 1, an enzyme implicated in synaptic vesicle recycling^{11,30}. The function of synaptojanin is tightly interconnected with that of dynamin, suggesting a mechanism that couples endocytic vesicle fission and phosphoinositide dephosphorylation¹¹. In budding yeast, conditional disruption of synaptojanin function results in ectopic accumulation of PtdIns(4,5)P₂ on internal membranes⁵⁵.

Phagocytosis and pathogen entry

Phagocytosis (that is, the engulfment of large particles ≥0.5 μm in diameter), begins with the activation of cell surface receptors by ligands that are present on the foreign particle⁵⁶. This event triggers

the local polymerization of actin, leading to the formation of cell protrusions (pseudopods) that engulf and ‘seal’ the particle. This process requires PtdIns(4,5)P₂ and local production of PtdIns(3,4,5)P₃, which stimulates actin nucleation. On ‘sealing’ however, catabolism of PtdIns(4,5)P₂—mediated at least in part by PLC—and of PtdIns(3,4,5)P₃ is critical for the shedding of actin and fusion of phagosomes with the endosome/lysosome compartment⁵⁶.

A special form of phagocytosis is the internalization of pathogenic bacteria, which capitalize in several ways on phosphoinositide-mediated signalling to invade and thrive in host cells (Table 1)⁵⁷. In a notable example of this interrelationship, some bacteria deliver phosphoinositide phosphatases into host cell cytoplasm that then help reprogram cell function to facilitate invasion and/or to determine the appropriate fate of pathogen-containing vacuoles (Table 1)^{17,57–61}.

Golgi complex function

PtdIns(4)P is the predominant phosphoinositide in the Golgi-complex region and a deficiency of this lipid affects its structure and function. In yeast, mutations of the Golgi-localized PtdIns(4)kinase PIK1 result in an accumulation of abnormal membrane intermediates (‘Berkeley bodies’) with impairment of transport to the plasma membrane¹⁸. These defects are reminiscent of those produced by mutations of the GTPase ARF1⁶², which may reflect the property of PtdIns(4)P to function as a co-receptor with ARF1 in the recruitment onto *trans*-Golgi network membranes of cytosolic proteins such as the clathrin adaptor AP-1, the GGA proteins and FAPP1/2 (refs 23, 63). The mammalian PIK1 homologue, PtdIns(4)kinase type IIIβ, interacts with and is stimulated by ARF1, indicating that this GTPase can cooperate with PtdIns(4)P by multiple and synergistic mechanisms⁶³ (see also Fig. 3b). In addition, disruption of Golgi integrity is produced by the knockdown of another PtdIns(4)kinase, type IIα (ref. 25).

Spatial restriction of PtdIns(4)P to the Golgi complex is ensured by phosphatases, one of which is likely to be Sac1 (ref. 64). This protein is an ER- and Golgi-complex-associated phosphoinositide phosphatase, which may thus dephosphorylate PtdIns(4)P when retrograde traffic occurs from the Golgi complex to the ER⁶⁴. Mutant *sac1* yeast cells have high levels of PtdIns(4)P, including high levels at the plasma membrane⁶⁵, which may be an indirect result of the excessive build-up of PtdIns(4)P in early secretory compartments.

The presence of pools of PtdIns(4,5)P₂ in the Golgi complex has been suggested by the reported localization of PtdIns(4,5)P₂-metabolizing enzymes, primarily the disease-linked inositol(5)phosphatase,

Table 1 | Phosphoinositide-metabolizing enzymes that have been linked to human diseases.

Gene name	Chr.	Enzyme	Predominant substrate	Product	Disease	Refs
PIK3CA	3	Class I PI(3)kinase p110α	PtdIns(4,5)P ₂	PtdIns(3,4,5)P ₃	Cancer	74
PIK3C3	18	Class III PI(3)kinase hVPS34	PtdIns	PtdIns(3)P	Bipolar disorder	75
PIK4CA	22	PtdIns(4)kinase Type IIα	PtdIns	PtdIns(4)P	Bipolar disorder	75
PIPSK2A	10	PtdInsP kinase Type IIα	PtdIns(5)P	PtdIns(4,5)P ₂	Bipolar disorder	75
PIPSK3	2	PtdInsP kinase Type III (PIKfyve)	PtdIns(3)P	PtdIns(3,5)P ₂	Francois-Neetens fleck cornea dystrophy	72
			PtdIns?	PtdIns(5)P?		
MTM1	X	3-phosphatase myotubularin	PtdIns(3)P	PtdIns	Myotubular myopathy	70
			PtdIns(3,5)P ₂	PtdIns(5)P		
MTMR2	11	3-phosphatase myotubularin-related protein	PtdIns(3)P	PtdIns	Charcot-Marie-Tooth type 4B1	71
			PtdIns(3,5)P ₂	PtdIns(5)P		
PTEN	10	3-phosphatase PTEN	PtdIns(3,4,5)P ₃	PtdIns(4,5)P ₂	Cancer	34, 35
INPPL1	11	5-phosphatase SHIP2	PtdIns(3,4,5)P ₃	PtdIns(3,4)P ₂	Type 2 diabetes	
OCRL1	X	5-phosphatase OCRL	PtdIns(4,5)P ₂	PtdIns(4)P	Oculocerebrorenal syndrome of Lowe	66
SYNJ1	21	polyphosphoinositide phosphatase synaptojanin 1	PtdIns(4,5)P ₂ Other phosphoinositides	PtdIns(4)P	Bipolar disorder, Down syndrome?	75
Bacterial gene		Enzyme	Predominant substrate	Product	Pathogen	Refs
BopB		4-phosphatase	PtdIns(4,5)P ₂	PtdIns(5)P	<i>Burkholderia pseudomallei</i>	61
IpgD		4-phosphatase	PtdIns(4,5)P ₂	PtdIns(5)P	<i>Shigella flexneri</i>	57, 58
SapM		3-phosphatase	PtdIns(3)P	PtdIns	<i>Mycobacterium tuberculosis</i>	60
SigD/SopB		polyphosphoinositide phosphatase	PtdIns(4,5)P ₂ Other phosphoinositides	PtdIns(5)P	<i>Salmonella enterica</i> /Dublin	57, 59

Refs 74, 75 are additional references for Table 1. Chr., chromosome. Bold, established genetic link; plain, putative link.

OCRL, in this region (Table 1)^{17,63,66}. The functional significance of PtdIns(4,5)P₂ in the Golgi complex remains unclear.

Endosome dynamics

PtdIns(3)P and PtdIns(3,5)P₂ have a crucial role in the biology of endosomes—a heterogeneous system of membranes that function as crossroads for traffic from and to the plasma membrane, the Golgi complex and the lysosomes¹⁸. PtdIns(3)P is a major determinant of early endosome membrane identity, and participates in nearly all aspects of endosomal function, such as membrane tethering and fusion, interaction with the cytoskeleton, signalling and motility. PtdIns(3)P-binding modules, primarily FYVE and PX domains, are present in a variety of endosomal proteins, such as EEA1, Hrs and SARA^{15,16,21,67}. These proteins are generally localized to distinct subdomains of the endosomal system⁶⁸, suggesting that PtdIns(3)P binding confers a first layer of specificity to their subcellular targeting and that other interactions fine-tune their precise localization. Several PtdIns(3)P-binding proteins also bind Rab5, an early endosome-associated GTPase, thus providing a powerful and highly specific coincidence detection mechanism for the recruitment of such proteins to early endosomes. Rab5, in turn, regulates PtdIns(3)P synthesis through two distinct pathways. The majority of this lipid is produced from PtdIns by the class III PI(3)kinase VPS34, a Rab5 effector^{10,18,68}. An additional pool of PtdIns(3)P is generated along the endocytic pathway through the sequential dephosphorylation of PtdIns(3,4,5)P₃ by inositol phosphatases specific for the 5 and 4 positions, respectively. The latter pathway, which is upregulated by growth factor stimulation, coordinates the internalization of PtdIns(3,4,5)P₃-enriched membranes with the modification of their phosphoinositide content to match the property of endosomal membranes²⁹.

PtdIns(3)P can undergo dephosphorylation at the endosomal surface, degradation within the lumen of multivesicular bodies, or phosphorylation to PtdIns(3,5)P₂ by the Fab1p/PIKfyve kinase^{18,69}. A main function of PtdIns(3,5)P₂ may be to trigger budding of vesicles from late endosomes and many of the effects of Fab1p/PIKfyve disruption (for example, swelling of the yeast vacuole and of late endosomes in mammalian cells) may be explained by defective membrane recycling from this compartment. Dephosphorylation of PtdIns(3)P and PtdIns(3,5)P₂ is achieved by members of the myotubularin³⁴ and Sac1 families⁶⁴.

Human genetic diseases have been mapped to genes encoding myotubularin family members^{70,71} and PIKfyve⁷², although the precise role of phosphoinositide metabolism imbalance in the phenotypic manifestations of these diseases remains poorly understood (Table 1).

Functions at other intracellular locations

Although PtdIns is synthesized in the ER, the presence of its phosphorylated derivatives in this compartment remains very much an open question. Phosphoinositide phosphatases located on this organelle, such as Sac1 and SKIP, may act primarily to dephosphorylate phosphoinositides inappropriately delivered to its membranes. Alternatively, they may act in *trans* on other membranes. The potential role of phosphoinositides in mitochondria remains unexplored, although enzymes, such as a synaptojanin splice variant, have been reported there^{19,63}. As mentioned above, phosphoinositides and phosphoinositide-binding proteins have been described in non-membranous locations in the nucleus²⁰, which may reflect the presence of proteins with hydrophobic pockets that are able to accommodate their fatty acid tails. Consistent with these observations, major phosphoinositide-metabolizing enzymes are found in the nucleus, and there is evidence for a role of phosphoinositides in pre-messenger-RNA splicing, chromatin remodelling and gene transcription²⁰.

Concluding remarks

As is clear from this overview, phosphoinositides occupy a general and central place in the field of cell signalling and regulation. A challenge for the future will be to elucidate how signalling effects

elicited by their rapid metabolic changes are coordinated with their constitutive functions in a variety of cell processes. Surprisingly, many key phosphoinositide metabolizing enzymes remain poorly characterized and the function of only a small subset of them has been genetically explored, at least in higher organisms. A central question is whether the continuous turnover of each phosphoinositide (referred to in some cases as a 'futile' cycle) primarily serves to allow for rapid changes and a fine regulation of their levels, or whether turnover is intimately connected to their mechanisms of action.

An important issue in phosphoinositide signalling is to what extent the levels of 'free' phosphoinositides are controlled by phosphoinositide-sequestering proteins (for example, 'pipmodulins')^{45,73}. At the plasma membrane, pipmodulins, such as MARCKS, are thought to buffer PtdIns(4,5)P₂ and to release it after appropriate stimuli (for example, protein phosphorylation), thus potentially contributing another layer of regulation⁷³.

One priority in phosphoinositide research is the further development of methods for their efficient detection and for the discrimination of phosphoinositides with different fatty acid compositions. Genetically encoded, fluorescent, phosphoinositide-binding modules are now commonly used to image phosphoinositides in living cells, but the presence in these modules of additional binding sites for cellular proteins (and/or other lipids), or the masking of phosphoinositides by endogenous proteins, may bias the results. The extent to which specific phosphoinositide pools are invisible to these probes is a highly debated issue, because the locations of some phosphoinositide-metabolizing enzymes seem to be in disagreement with the predominant location of the substrates and/or products revealed by such probes. Whether pools of phosphoinositides at locations that are not revealed by these probes do exist and are physiologically important remains an open question.

Finally, the involvement of several key phosphoinositide metabolic reactions in disease makes this area of research highly relevant to medicine (see also Table 1). Further elucidation of the role of phosphoinositides in cell biology and organismal physiology is expected to provide new potential targets for pharmacological interventions in major human diseases, including cancer and diabetes.

- Hokin, L. E. Receptors and phosphoinositide-generated second messengers. *Annu. Rev. Biochem.* **54**, 205–235 (1985).
- Berridge, M. J. & Irvine, R. F. Inositol phosphates and cell signalling. *Nature* **341**, 197–205 (1989).
- Lassing, I. & Lindberg, U. Specific interaction between phosphatidylinositol 4,5-bisphosphate and profilactin. *Nature* **314**, 472–474 (1985).
- Ma, L., Cantley, L. C., Janmey, P. A. & Kirschner, M. W. Corequirement of specific phosphoinositides and small GTP-binding protein Cdc42 in inducing actin assembly in *Xenopus* egg extracts. *J. Cell Biol.* **140**, 1125–1136 (1998).
- Yin, H. L. & Janmey, P. A. Phosphoinositide regulation of the actin cytoskeleton. *Annu. Rev. Physiol.* **65**, 761–789 (2003).
- Whitman, M., Downes, C. P., Keeler, M., Keller, T. & Cantley, L. Type I phosphatidylinositol kinase makes a novel inositol phospholipid, phosphatidylinositol-3-phosphate. *Nature* **332**, 644–646 (1988).
- Traynor-Kaplan, A. E., Harris, A. L., Thompson, B. L., Taylor, P. & Sklar, L. A. An inositol tetrakisphosphate-containing phospholipid in activated neutrophils. *Nature* **334**, 353–356 (1988).
- Auger, K. R. *et al.* PDGF-dependent tyrosine phosphorylation stimulates production of novel polyphosphoinositides in intact cells. *Cell* **57**, 167–175 (1989).
- Eberhard, D. A., Cooper, C. L., Low, M. G. & Holz, R. W. Evidence that the inositol phospholipids are necessary for exocytosis. Loss of inositol phospholipids and inhibition of secretion in permeabilized cells caused by a bacterial phospholipase C and removal of ATP. *Biochem. J.* **268**, 15–25 (1990).
- Schu, P. V. *et al.* Phosphatidylinositol 3-kinase encoded by yeast VPS34 gene essential for protein sorting. *Science* **260**, 88–91 (1993).
- McPherson, P. S. *et al.* A presynaptic inositol-5-phosphatase. *Nature* **379**, 353–357 (1996).
- Martin, T. F. Phosphoinositide lipids as signaling molecules: common themes for signal transduction, cytoskeletal regulation, and membrane trafficking. *Annu. Rev. Cell Dev. Biol.* **14**, 231–264 (1998).
- Dove, S. K. *et al.* Osmotic stress activates phosphatidylinositol-3,5-bisphosphate synthesis. *Nature* **390**, 187–192 (1997).
- Rameh, L. E., Tolias, K. F., Duckworth, B. C. & Cantley, L. C. A new pathway for synthesis of phosphatidylinositol-4,5-bisphosphate. *Nature* **390**, 192–196 (1997).
- Lemmon, M. A. Phosphoinositide recognition domains. *Traffic* **4**, 201–213 (2003).

16. Balla, T. Inositol-lipid binding motifs: signal integrators through protein–lipid and protein–protein interactions. *J. Cell Sci.* **118**, 2093–2104 (2005).
17. Majerus, P. W., Kisseleva, M. V. & Norris, F. A. The role of phosphatases in inositol signaling reactions. *J. Biol. Chem.* **274**, 10669–10672 (1999).
18. Odorizzi, G., Babst, M. & Emr, S. D. Phosphoinositide signaling and the regulation of membrane trafficking in yeast. *Trends Biochem. Sci.* **25**, 229–235 (2000).
19. Roth, M. G. Phosphoinositides in constitutive membrane traffic. *Physiol. Rev.* **84**, 699–730 (2004).
20. Hammond, G., Thomas, C. L. & Schiavo, G. Nuclear phosphoinositides and their functions. *Curr. Top. Microbiol. Immunol.* **282**, 177–206 (2004).
21. Hurley, J. H. & Meyer, T. Subcellular targeting by membrane lipids. *Curr. Opin. Cell Biol.* **13**, 146–152 (2001).
22. Wenk, M. R. & De Camilli, P. Protein–lipid interactions and phosphoinositide metabolism in membrane traffic: insights from vesicle recycling in nerve terminals. *Proc. Natl Acad. Sci. USA* **101**, 8262–8269 (2004).
23. Behnia, R. & Munro, S. Organelle identity and the signposts for membrane traffic. *Nature* **438**, 597–604 (2005).
24. Gaidarov, I. & Keen, J. H. Phosphoinositide–AP-2 interactions required for targeting to plasma membrane clathrin-coated pits. *J. Cell Biol.* **146**, 755–764 (1999).
25. Wang, Y. J. *et al.* Phosphatidylinositol 4 phosphate regulates targeting of clathrin adaptor AP-1 complexes to the Golgi. *Cell* **114**, 299–310 (2003).
26. Honing, S. *et al.* Phosphatidylinositol-(4,5)-bisphosphate regulates sorting signal recognition by the clathrin-associated adaptor complex AP2. *Mol. Cell* **18**, 519–531 (2005).
27. Owen, D. J., Collins, B. M. & Evans, P. R. Adaptors for clathrin coats: structure and function. *Annu. Rev. Cell Dev. Biol.* **20**, 153–191 (2004).
28. Honda, A. *et al.* Phosphatidylinositol 4-phosphate 5-kinase α is a downstream effector of the small G protein ARF6 in membrane ruffle formation. *Cell* **99**, 521–532 (1999).
29. Shin, H. W. *et al.* An enzymatic cascade of Rab5 effectors regulates phosphoinositide turnover in the endocytic pathway. *J. Cell Biol.* **170**, 607–618 (2005).
30. Cremona, O. *et al.* Essential role of phosphoinositide metabolism in synaptic vesicle recycling. *Cell* **99**, 179–188 (1999).
31. Cantley, L. C. The phosphoinositide 3-kinase pathway. *Science* **296**, 1655–1657 (2002).
32. Czech, M. P. Dynamics of phosphoinositides in membrane retrieval and insertion. *Annu. Rev. Physiol.* **65**, 791–815 (2003).
33. Katso, R. *et al.* Cellular function of phosphoinositide 3-kinases: implications for development, homeostasis, and cancer. *Annu. Rev. Cell Dev. Biol.* **17**, 615–675 (2001).
34. Wishart, M. J. & Dixon, J. E. PTEN and myotubularin phosphatases: from 3-phosphoinositide dephosphorylation to disease. *Trends Cell Biol.* **12**, 579–585 (2002).
35. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1947 (1997).
36. Hilgemann, D. W., Feng, S. & Nasuhoglu, C. The complex and intriguing lives of PIP₂ with ion channels and transporters. *Sci. STKE* **2001**, RE19, doi:10.1126/stke.2001.111.re19 (2001).
37. Suh, B. C. & Hille, B. Regulation of ion channels by phosphatidylinositol 4,5-bisphosphate. *Curr. Opin. Neurobiol.* **15**, 370–378 (2005).
38. Pollard, T. D. & Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* **112**, 453–465 (2003).
39. Rohatgi, R., Ho, H. Y. & Kirschner, M. W. Mechanism of N-WASP activation by CDC42 and phosphatidylinositol 4,5-bisphosphate. *J. Cell Biol.* **150**, 1299–1310 (2000).
40. Ho, H. Y. *et al.* Toca-1 mediates Cdc42-dependent actin nucleation by activating the N-WASP–WIP complex. *Cell* **118**, 203–216 (2004).
41. Di Paolo, G. *et al.* Recruitment and regulation of phosphatidylinositol phosphate kinase type 1 γ by the FERM domain of talin. *Nature* **420**, 85–89 (2002).
42. Ling, K. *et al.* Type 1 γ phosphatidylinositol phosphate kinase targets and regulates focal adhesions. *Nature* **420**, 89–93 (2002).
43. Oikawa, T. *et al.* PtdIns(3,4,5)P₃ binding is necessary for WAVE2-induced formation of lamellipodia. *Nature Cell Biol.* **6**, 420–426 (2004).
44. Iijima, M., Huang, Y. E. & Devreotes, P. Temporal and spatial regulation of chemotaxis. *Dev. Cell* **3**, 469–478 (2002).
45. Golub, T. & Caroni, P. PI(4,5)P₂-dependent microdomain assemblies capture microtubules to promote and control leading edge motility. *J. Cell Biol.* **169**, 151–165 (2005).
46. Klopstein, D. R. & Vale, R. D. The lipid binding pleckstrin homology domain in UNC-104 kinesin is necessary for synaptic vesicle transport in *Caenorhabditis elegans*. *Mol. Biol. Cell* **15**, 3729–3739 (2004).
47. Bai, J., Tucker, W. C. & Chapman, E. R. PIP₂ increases the speed of response of synaptotagmin and steers its membrane-penetration activity toward the plasma membrane. *Nature Struct. Mol. Biol.* **11**, 36–44 (2004).
48. Rothman, J. E. The protein machinery of vesicle budding and fusion. *Protein Sci.* **5**, 185–194 (1996).
49. Milosevic, I. *et al.* Plasmalemmal phosphatidylinositol-4,5-bisphosphate level regulates the releasable vesicle pool size in chromaffin cells. *J. Neurosci.* **25**, 2557–2565 (2005).
50. Gong, L. W. *et al.* Phosphatidylinositol phosphate kinase type 1 γ regulates dynamics of large dense-core vesicle fusion. *Proc. Natl Acad. Sci. USA* **102**, 5204–5209 (2005).
51. Di Paolo, G. *et al.* Impaired PtdIns(4,5)P₂ synthesis in nerve terminals produces defects in synaptic vesicle trafficking. *Nature* **431**, 415–422 (2004).
52. Lackner, M. R., Nurrish, S. J. & Kaplan, J. M. Facilitation of synaptic transmission by EGL-30 Gq α and EGL-8 PLC β : DAG binding to UNC-13 is required to stimulate acetylcholine release. *Neuron* **24**, 335–346 (1999).
53. Rhee, J. S. *et al.* β phorbol ester- and diacylglycerol-induced augmentation of transmitter release is mediated by Munc13s and not by PKCs. *Cell* **108**, 121–133 (2002).
54. Engqvist-Goldstein, A. E. & Drubin, D. G. Actin assembly and endocytosis: from yeast to mammals. *Annu. Rev. Cell Dev. Biol.* **19**, 287–332 (2003).
55. Stefan, C. J., Audhya, A. & Emr, S. D. The yeast synaptojanin-like proteins control the cellular distribution of phosphatidylinositol (4,5)-bisphosphate. *Mol. Biol. Cell* **13**, 542–557 (2002).
56. Botelho, R. J., Scott, C. C. & Grinstein, S. Phosphoinositide involvement in phagocytosis and phagosome maturation. *Curr. Top. Microbiol. Immunol.* **282**, 1–30 (2004).
57. Pizarro-Cerda, J. & Cossart, P. Subversion of phosphoinositide metabolism by intracellular bacterial pathogens. *Nature Cell Biol.* **6**, 1026–1033 (2004).
58. Niebuhr, K. *et al.* Conversion of PtdIns(4,5)P₂ into PtdIns(5)P by the *S. flexneri* effector IpgD reorganizes host cell morphology. *EMBO J.* **21**, 5069–5078 (2002).
59. Terebiznik, M. R. *et al.* Elimination of host cell PtdIns(4,5)P₂ by bacterial SigD promotes membrane fission during invasion by *Salmonella*. *Nature Cell Biol.* **4**, 766–773 (2002).
60. Vergne, I. *et al.* Mechanism of phagolysosome biogenesis block by viable *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **102**, 4033–4038 (2005).
61. Ungewickell, A. *et al.* The identification and characterization of two phosphatidylinositol-4,5-bisphosphate 4-phosphatases. *Proc. Natl Acad. Sci. USA* **102**, 18854–18859 (2005).
62. Gaynor, E. C., Chen, C. Y., Emr, S. D. & Graham, T. R. ARF is required for maintenance of yeast Golgi and endosome structure and function. *Mol. Biol. Cell* **9**, 653–670 (1998).
63. De Matteis, M. A., Di Campli, A. & Godi, A. The role of the phosphoinositides at the Golgi complex. *Biochim. Biophys. Acta* **1744**, 396–405 (2005).
64. Guo, S., Stolz, L. E., Lemrow, S. M. & York, J. D. SAC1-like domains of yeast SAC1, INP52, and INP53 and of human synaptojanin encode polyphosphoinositide phosphatases. *J. Biol. Chem.* **274**, 12990–12995 (1999).
65. Roy, A. & Levine, T. P. Multiple pools of phosphatidylinositol 4-phosphate detected using the pleckstrin homology domain of Osh2p. *J. Biol. Chem.* **279**, 44683–44689 (2004).
66. Attree, O. *et al.* The Lowe's oculocerebrorenal syndrome gene encodes a protein highly homologous to inositol polyphosphate-5-phosphatase. *Nature* **358**, 239–242 (1992).
67. Birkeland, H. C. & Stenmark, H. Protein targeting to endosomes and phagosomes via FYVE and PX domains. *Curr. Top. Microbiol. Immunol.* **282**, 89–115 (2004).
68. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* **2**, 107–117 (2001).
69. Michell, R. H., Heath, V. L., Lemmon, M. A. & Dove, S. K. Phosphatidylinositol 3,5-bisphosphate: metabolism and cellular functions. *Trends Biochem. Sci.* **52**–63 (2005).
70. Laporte, J. *et al.* A gene mutated in X-linked myotubular myopathy defines a new putative tyrosine phosphatase family conserved in yeast. *Nature Genet.* **13**, 175–182 (1996).
71. Bolino, A. *et al.* Charcot-Marie-Tooth type 4B is caused by mutations in the gene encoding myotubularin-related protein-2. *Nature Genet.* **25**, 17–19 (2000).
72. Li, S. *et al.* Mutations in PIP5K3 are associated with Francois-Neetens mouchette fleck corneal dystrophy. *Am. J. Hum. Genet.* **77**, 54–63 (2005).
73. McLaughlin, S. & Murray, D. Plasma membrane phosphoinositide organization by protein electrostatics. *Nature* **438**, 605–611 (2005).
74. Bader, A. G., Kang, S., Zhao, L. & Vogt, P. K. Oncogenic PI3K deregulates transcription and translation. *Nature Rev. Cancer* **5**, 921–929 (2005).
75. Halstead, J. R., Jalink, K. & Divecha, N. An emerging role for PtdIns(4,5)P₂-mediated signalling in human disease. *Trends Pharmacol. Sci.* **26**, 654–660 (2005).

Acknowledgements We thank N. Borgese, V. Haucke and O. Cremona for critical reading of the manuscript. We also thank B. Chang for providing the fluorescence images in Fig. 1 and for her comments on the manuscript. DNA constructs were gifts from H. Stenmark, T. Meyer and A. De Matteis. We apologize to all the scientists whose original studies and reviews were not quoted in our manuscript owing to space limitations. G.D.P. is funded by grants from the National Institute of Health. P.D.C. is funded by the HHMI and by grants from the National Institute of Health, the Yale Center for Genomics and Proteomics, the Yale/NIDA Neuroproteomics Center and the G. Harold and Leila Y. Mathers Charitable Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to P.D.C. (pietro.decamilli@yale.edu) and G.D.P. (gd2175@columbia.edu).

ARTICLES

Role of Bax and Bak in mitochondrial morphogenesis

Mariusz Karbowski^{1*†}, Kristi L. Norris^{1*}, Megan M. Cleland¹, Seon-Yong Jeong^{1†} & Richard J. Youle¹

Bcl-2 family proteins are potent regulators of programmed cell death. Although their intracellular localization to mitochondria and the endoplasmic reticulum has focused research on these organelles, how they function remains unknown. Two members of the Bcl-2 family, Bax and Bak, change intracellular location early in the promotion of apoptosis to concentrate in focal clusters at sites of mitochondrial division. Here we report that in healthy cells Bax or Bak is required for normal fusion of mitochondria into elongated tubules. Bax seems to induce mitochondrial fusion by activating assembly of the large GTPase Mfn2 and changing its submitochondrial distribution and membrane mobility—properties that correlate with different GTP-bound states of Mfn2. Our results show that Bax and Bak regulate mitochondrial dynamics in healthy cells and indicate that Bcl-2 family members may also regulate apoptosis through organelle morphogenesis machineries.

Programmed cell death is essential in multicellular organisms for development, tissue turnover and host defence. In mammalian cells, mitochondria have a central role in apoptosis that is regulated by members of the Bcl-2 family¹. Activation of either Bax or Bak, two homologous pro-apoptotic members of the Bcl-2 family, is associated with changes in their conformation that are postulated to induce permeabilization of the outer mitochondrial membrane (OMM)². Bax and Bak double knockout (Bax/Bak DKO) cells show resistance to apoptosis induced by various treatments, demonstrating the crucial role of these proteins in the regulation of cell death³.

During apoptosis, mitochondria fragment into smaller units. Simultaneous with this mitochondrial fragmentation, Bax and Bak change their intracellular locations to concentrate into foci at sites of mitochondrial scission, where they colocalize with mitofusin 2 (Mfn2) and dynamin-related protein 1 (Drp1)⁴, two proteins that mediate mitochondrial fusion and fission, respectively. We therefore examined the possibility that Bax and Bak could regulate mitochondrial morphology in healthy cells.

Mitochondrial morphology depends on Bax or Bak

Mitochondrial shape and size were examined in six independent cell lines or explants from mice lacking expression of both Bax and Bak. We compared primary and transformed mouse embryonic fibroblasts (Bax/Bak DKO MEFs)³ and baby mouse kidney cell lines (Bax/Bak DKO BMKs)⁵ from Bax/Bak DKO mice with lineage-matched cell lines from wild-type mouse strains. Cells were labelled with the mitochondria-specific probe Mitotracker Red (data not shown) or by immunostaining the mitochondrial marker cytochrome *c* (Fig. 1a, b). In $68.1 \pm 6.7\%$ (mean $\pm$ s.d.) of transformed Bax/Bak DKO MEFs, more than $\sim 90\%$ of mitochondria in a cell showed a short, fragmented morphology and only occasionally contained elongated tubules (Fig. 1b, c). By contrast, most single Bax or single Bak knockout cells ($94.8 \pm 1.5\%$ and $80.0 \pm 4.7\%$ of Bax KO and Bak KO, respectively) had elongated tubular mitochondria resembling those in $80.9 \pm 8.8\%$ of wild-type MEFs from the same mouse mixed background 129/CD1 (ref. 3 and Fig. 1c), in

addition to those in MEFs previously described⁶ from a 129/SvEv background ($98.6 \pm 0.7\%$ had long branching mitochondria). Similar results were seen in primary MEF explants from two Bax/Bak DKO mice as compared with cells from one wild-type mouse and one *Bax*^{-/-}*Bak*^{+/+} mouse (Supplementary Fig. S1). In addition, only $11 \pm 4.2\%$ and $8 \pm 2.8\%$ of the cells had tubular, elongated mitochondria in two independent Bax/Bak DKO BMK cell lines, in marked contrast to two independent wild-type BMK cell lines, which had elongated tubular mitochondria ($83 \pm 4.2\%$ and $40 \pm 2.8\%$; Supplementary Fig. S1).

The average length of a mitochondrion was substantially shorter in MEFs from three different Bax/Bak DKO mice ($3.3\text{--}4.4\ \mu\text{m}$), as compared with MEFs from three different control mice ($8.4\text{--}11.1\ \mu\text{m}$); and in Bax/Bak DKO BMK cells ($2.8 \pm 1.7\ \mu\text{m}$), as compared with wild-type BMKs ($8.0 \pm 2.1\ \mu\text{m}$; Fig. 1g). The extent of mitochondrial fragmentation in the Bax/Bak DKO cell lines was comparable to that induced by the knockout of Mfn1 or Mfn2 ($2.4 \pm 0.6\ \mu\text{m}$ in Mfn2 KO versus $11.5 \pm 5.5\ \mu\text{m}$ in 129/SvEv wild-type MEFs, Fig. 1g; ref. 6). Bone-marrow-derived cells from a Bax/Bak DKO mouse also showed markedly fragmented mitochondria (Fig. 1d, f), and permanent re-expression of Bak into this cell line⁷ induced considerable mitochondrial elongation in healthy cells (Fig. 1e, f). Moreover, knock down of Bax and Bak by RNA interference (RNAi) in HeLa cells led to an overall decrease in mitochondrial size and network complexity (Supplementary Fig. S2), further confirming the requirement of Bax and Bak for mitochondrial morphogenesis.

Mitochondrial network complexity in cells lacking Bax and Bak was also evaluated by fluorescence recovery after photobleaching (FRAP; Fig. 1h), a technique that, by assessing the depth of bleach and rate of redistribution of mitochondrial matrix-localized yellow fluorescence protein (mito-YFP) into regions of irreversible fluorophore photobleach, provides a semiquantitative measure of the volume of single units of the mitochondrial network⁸. Fast but minor refilling of bleached regions occurred in Bax/Bak DKO cells (to $40.15 \pm 6.3\%$ of initial fluorescence; mean $\pm$ s.d.), comparable

¹Biochemistry Section, SNB, NINDS, NIH, Bethesda, Maryland 20892, USA. [†]Present addresses: University of Maryland Biotechnology Institute, Medical Biotechnology Center, UMBI Building, 725 West Lombard Street, Baltimore, Maryland 21201, USA (M.K.); Department of Medical Genetics, School of Medicine, Ajou University, Suwon 443-721, Korea (S.-Y.J.).

*These authors contributed equally to this work.

to that in Mfn2 KO cells (to $44.9 \pm 7.4\%$ of initial fluorescence). By contrast, refilling of bleached areas in wild-type MEFs or single Bak KO MEFs was more prominent (to $59.41 \pm 11.16\%$ and $58.2 \pm 6.2\%$ of initial fluorescence, respectively), indicating that diffusion of the mito-YFP occurred from a greater distance and volume in wild-type MEFs than in Bax/Bak DKO or Mfn2 KO cells (Fig. 1h). Thus, the shorter mitochondria in Bax/Bak DKO cells are discontinuous, separate units.

Bax overexpression elongates mitochondria

The cytomegalovirus protein, viral mitochondria-associated inhibitor of apoptosis (vMIA), prevents apoptosis by sequestering and inactivating Bax^{9,10} and Bak (Supplementary Fig. S3). As reported in primary human fibroblasts¹¹, vMIA expression in DU145 cells (Fig. 2b, compared with control cells in Fig. 2a) and HCT116 Bax^{-/-} cells (data not shown) caused the mitochondria to fragment into shorter units. We reasoned that binding of Bax and Bak and inhibition of their mitochondrial elongation activity by vMIA may cause this mitochondrial fragmentation. Ectopic expression of Bax reversed the vMIA-induced mitochondrial fragmentation, as assessed visually and by FRAP in DU145 cells and in HCT116 cells (Fig. 2c, e, f and data not shown). Bax lacking the BH3 domain (Fig. 2d–f), an anti-apoptotic and a BH3-only member of the Bcl-2 family (Supplementary Fig. S4), did not reverse vMIA-induced mitochondrial fragmentation. These data indicate that specific activities of Bax and Bak, which might be linked to their apoptosis-related functions, are required for the formation of balanced mitochondrial networks.

Mitochondrial fission and fusion machinery

On inhibition of mitochondrial fission by loss of Drp1 activity, mitochondria become unusually interconnected owing to unbalanced

fusion^{12,13}. Inactivation of Drp1 with the dominant-negative inhibitor Drp1-K38A¹², however, did not cause mitochondrial elongation in Bax/Bak DKO MEFs (Supplementary Fig. S5H) or in vMIA-expressing DU145 cells (Supplementary Figs S5F–H and S6D), HCT116 Bax^{-/-} cells (data not shown) or Cos-7 cells (Supplementary Fig. S5A–E), indicating that the fusion step of mitochondrial morphogenesis requires Bax or Bak activity. Use of Drp1 RNAi¹⁴ confirmed that inhibition of mitochondrial fission caused less elongation of mitochondria in Bax/Bak DKO cells than in Mfn2 KO cells (Supplementary Fig. S7).

Further indicating a role of Bax/Bak in mitochondrial fusion, ectopic overexpression of Mfn2 overcame the fusion deficit in Bax/Bak DKO cells (Fig. 3), and this was dependent on the Mfn2 carboxy-terminal coiled-coil domain (Supplementary Fig. S11D). In addition, ectopic expression of both Bax and Mfn2 in Bax/Bak DKO cells induced greater mitochondrial interconnectivity than Mfn2 alone (Fig. 3c), suggesting that Bax functions in mitochondrial morphogenesis by regulating Mfn2.

Mfn2 GTPase activity and Bax/Bak expression

The subcellular location, expression level and mobility of fission proteins Drp1 and Fis1 were not detectably different with and without Bax/Bak expression (Supplementary Fig. S8 and data not shown). Although the expression level of endogenous Mfn2 was also the same (Fig. 4f and Supplementary Fig. S8), the distribution of Mfn2 in the OMM in wild-type MEFs differed from that in Bax/Bak DKO MEFs. In wild-type MEFs, Mfn2 localized partially around the OMM, but primarily concentrated into foci at sites of mitochondrial fusion and fission^{4,15} (Fig. 4a and Supplementary Fig. S9), and the relative amounts of Mfn2 in the different locations could be quantified by analysis of confocal images (Fig. 4d and Table 1). In Bax/Bak DKO MEFs, however, Mfn2 coated the OMM more evenly and had

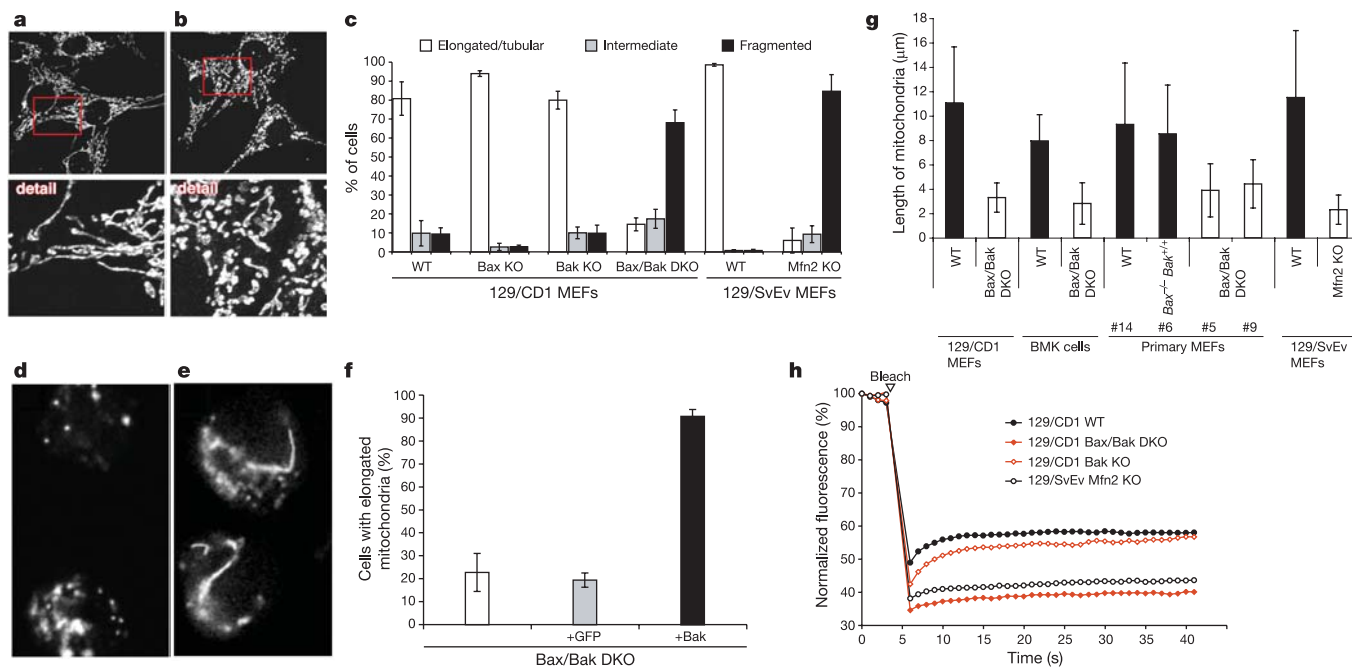


Figure 1 | Mitochondrial morphology defects in Bax/Bak DKO cells.

a, b, Wild-type (**a**) and Bax/Bak DKO (**b**) MEFs stained with antibodies to cytochrome *c* (ref. 19) were analysed by confocal microscopy. **c**, Mitochondrial morphology was scored in the cell types indicated. Data represent the mean $\pm$ s.d. of three independent experiments, each with >100 cells counted per cell type. **d–f**, IL-3-dependent bone-marrow-derived cells from DKO mice (**d**) and DKO mice permanently re-expressing Bak (**e**; clone Bak-8) stained with Mitotracker were analysed by confocal microscopy and compared with GFP-transfected IL-3-dependent bone-marrow-derived cells from Bax/Bak DKO mice (**f**). Cells with

predominantly elongated mitochondria were scored; data represent the mean $\pm$ s.d. of three experiments, each with 50 cells per condition.

g, Mitochondrial length was measured in the indicated cell types (see Supplementary Methods). Data represent the mean $\pm$ s.d. **h**, Mitochondrial network complexity in wild-type, Bax/Bak DKO, Bak KO and Mfn2 KO MEFs was measured by FRAP as described⁸. Differences in the mean fluorescence recovery between wild-type and Bax/Bak DKO MEFs or between wild-type and Mfn2 KO MEFs measured at 40 s were statistically significant at $P < 0.001$ (two-tailed *t*-test).

diminished foci (Fig. 4b, d, and Supplementary Fig. S9). Re-expression of Bax in Bax/Bak DKO MEFs significantly ($P = 0.01$) reconstituted focal localization of Mfn2 (Fig. 4c, d, and Supplementary Fig. S9). The ratio of Mfn2 in foci relative to that circumscribing the mitochondria increased more than 12-fold in the presence of either endogenous Bax/Bak or ectopic Bax (Fig. 4d).

We used FRAP to quantify Mfn2 membrane mobility in the presence or absence of Bax expression. Mfn2-YFP was significantly less mobile than three other proteins localizing to the OMM—OMP25 ($P < 0.005$), Fis1 ($P < 0.005$), or YFP fused to the membrane anchor of Bcl-x_L ($P < 0.005$) (Supplementary Fig. S10)—consistent with its more focal distribution. Deletion of the Mfn2 C-terminal coiled-coil domain (Mfn2(1–703)–YFP), which mediates Mfn2 complex formation¹⁶, led to increased membrane mobility (Fig. 4e) and shifted Mfn2 localization to circumscribe the mitochondria completely (Supplementary Fig. S11A–C). By contrast, the GTPase-inactive mutant Mfn2-K109T showed less mobility in the OMM than did wild-type Mfn2 (Fig. 4e), consistent with its increased focal mitochondrial distribution¹⁵ (Supplementary Fig. S11E, F).

Thus, various mutants of Mfn2 show a correlation between their degree of membrane mobility and location on the mitochondrial surface. The focal form seems to be considerably less mobile than the circumscribing population. Consistent with the shift of Mfn2 into foci along the OMM due to Bax/Bak expression, introducing Bax into Bax/Bak DKO MEFs caused a significant decrease in Mfn2 mobility ($P < 0.05$; Fig. 4e). However, Bax expression did not affect the mobility or distribution of Mfn2(1–703)–YFP ($P > 0.1$), Mfn2-K109T ($P > 0.1$) or a control protein in the OMM—the regulator of mitochondrial fission, Fis1 (Fig. 4e, Table 1 and Supplementary Fig. S11C). We therefore examined the mobility of endogenous Mfn2 using blue native gel electrophoresis (BN–PAGE) in the absence and

presence of Bax/Bak (Fig. 4f). Mfn2 migrated in a series of different size complexes, varying in molecular weight between 240 and 669 kDa, in 129/CD1 Bax/Bak DKO cells, but assembled into a single 440-kDa complex with Bax/Bak expression in wild-type MEFs. Ectopic expression of Bax in Bax/Bak DKO cells also shifted Mfn2 into 440-kDa complexes (data not shown). However, Opa1, another dynamin family protein required for mitochondrial fusion, did not show any change in complex formation, as assessed by BN–PAGE, between wild-type and Bax/Bak DKO MEFs (Supplementary Fig. S12), showing the specificity of Bax modulation of Mfn2 complex formation. Neither incubation of GDP nor incubation of GTP- γ S with the cell lysates affected the 440-kDa complex formation (Fig. 4f). Gel filtration of haemagglutinin-tagged Mfn2 also showed a change in Mfn2 complex size owing to the expression of ectopic Bax (Supplementary Fig. S13). These three changes in Mfn2—localization, membrane mobility and complex formation (Fig. 4)—that are regulated by Bax expression may reflect a mechanism of Bax/Bak-induced mitochondrial tubule maintenance. In contrast to the GTPase-inactive mutant Mfn2-K109T, a mutant of Mfn2 with reduced GTPase activity and increased GDP–GTP exchange to favour the GTP bound-form, Mfn2-G12V, has been shown to circumscribe the mitochondria with reduced focal distribution¹⁵. We confirmed that Mfn2-G12V primarily circumscribes the OMM and found that, unlike wild-type Mfn2, it is not shifted into foci by Bax (Table 1 and Supplementary Fig. S11G, H). Notably, Mfn2-G12V fully reconstituted mitochondrial fusion in Mfn2 KO MEFs but not in Bax/Bak DKO MEFs (Fig. 3d–g).

These data indicate that Bax or Bak is required for Mfn2-activity-dependent mitochondrial fusion. We therefore directly examined mitochondrial fusion in Bax/Bak DKO cells by using a quantitative assay based on the redistribution and fluorescence dilution of a mitochondrial matrix-targeted photoactivable green fluorescent protein (mito-PAGFP)¹⁷. Small regions of interest (ROIs) in wild-type

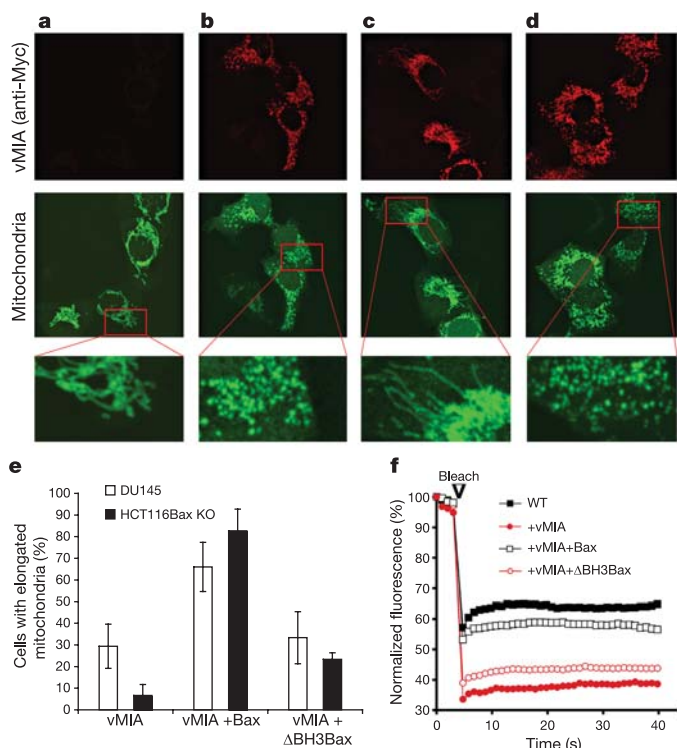


Figure 2 | vMIA-induced mitochondrial fragmentation is reversed by Bax overexpression. **a–e**, Mito-YFP labelled mitochondria (green) in DU145 cells (**a–c**) and HCT116 Bax KO cells (**e**) expressing the indicated constructs were imaged after anti-Myc immunostaining of vMIA (red) and scored for elongation. Data represent the mean $\pm$ s.d. of three experiments, each with 150 cells scored per condition. **f**, DU145 cells were analysed for mitochondrial contiguity by FRAP. Data represent an average of >30 measurements obtained in at least two experiments.

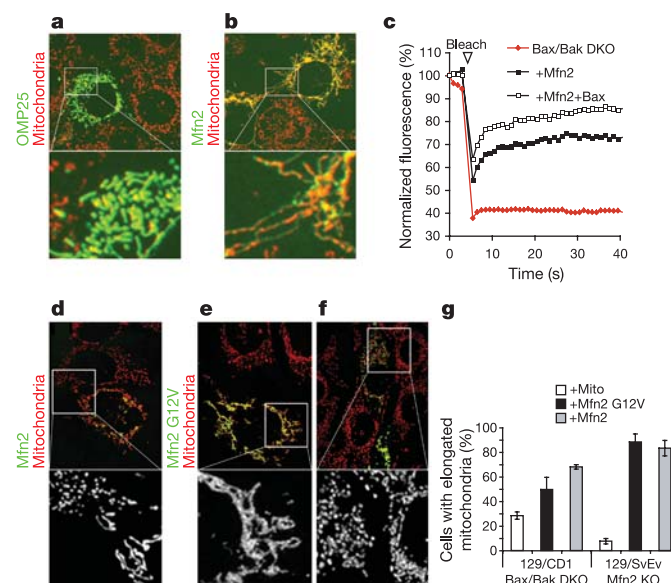


Figure 3 | Ectopic Mfn2 but not Mfn2-G12V reverses fragmentation owing to loss of Bax/Bak. **a, b**, Images of Bax/Bak DKO cells transfected with YFP–OMP25 (**a**; green) or with Mfn2–YFP (**b**; yellow) and stained for cytochrome *c* to visualize mitochondria (red). **c**, FRAP analysis of Bax/Bak DKO cells expressing mito-YFP alone or mito-YFP with Mfn2 or with Mfn2 plus Bax. **d–f**, Images of Mfn2 KO (**d, e**) and Bax/Bak DKO (**f**) MEFs transfected with Mfn2–YFP (**d**; green) or Mfn2-G12V–YFP (**e, f**; green) and immunostained for cytochrome *c* (**d–f**; red, top; white, bottom). **g**, Effects of Mfn2–YFP and Mfn2-G12V–YFP on mitochondrial morphology in Mfn2 KO cells and Bax/Bak DKO cells assessed by blind counting. Data represent the mean $\pm$ s.d. of two experiments, each with triplicates of 50 cells counted per condition.

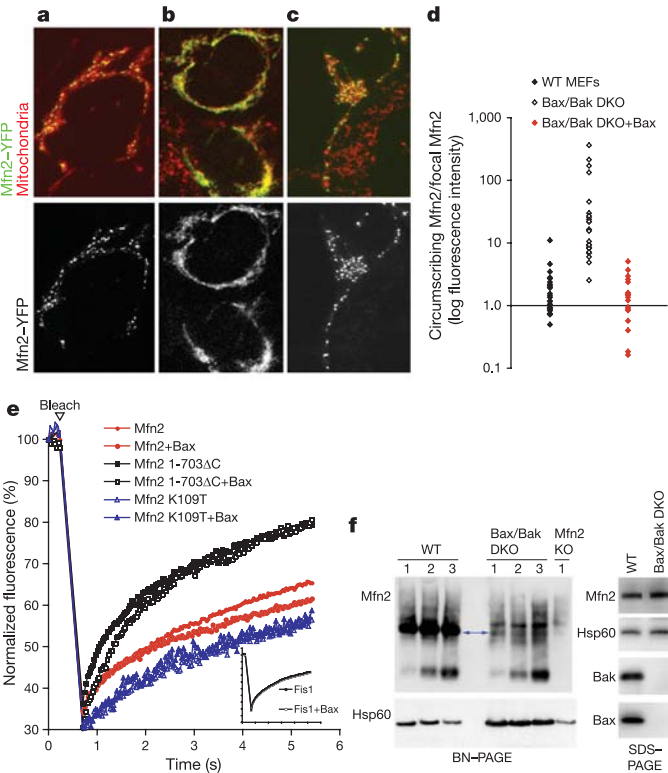


Figure 4 | Mfn2 complex formation, subcellular distribution and membrane mobility are altered by Bax expression. **a–c**, Wild-type (a) and Bax/Bak DKO (b, c) MEFs transiently transfected with Mfn2-YFP alone, and Bax/Bak DKO MEFs coexpressing Mfn2-YFP and Bax (c) were stained for cytochrome c and analysed by confocal microscopy for Mfn2 distribution (green or white) and mitochondria (red). **d**, Wild-type MEFs, Bax/Bak DKO MEFs transfected with Mfn2-YFP, and Bax/Bak DKO MEFs cotransfected with Mfn2-YFP and Bax were analysed for YFP distribution along the mitochondria (see Supplementary Methods). **e**, OMM mobility of Mfn2-YFP, Mfn2(1–703)-YFP, Mfn2-K109T-YFP and YFP-Fis1 (inset) expressed in Bax/Bak DKO MEFs and in the same cells cotransfected with Bax was measured by FRAP. Each line represents the mean of ≥ 30 measurements. **f**, Mfn2 complexes in wild-type, Bax/Bak DKO and Mfn2 KO MEFs were analysed by BN-PAGE¹⁹. Isolated mitochondria were incubated at 35 °C for 20 min in the absence (lane 1) or presence of 100 μ M GTP- γ S (lane 2) or 1 mM GDP (lane 3). Arrow indicates the 440-kDa band. SDS gels of whole-cell lysates show similar expression of Mfn2 with and without Bax/Bak expression.

Table 1 | Bax-altered distribution of Mfn2 and mutant Mfn2 along mitochondria

	Average area of circumscribing Mfn2-YFP per cell (a.u.)	P value ($\pm$ Bax)
Mfn2	55.9151	$P = 0.00226$
Mfn2+Bax	31.5136	
Mfn2-K109T	11.1433	$P = 0.34274$
Mfn2-K109T+Bax	11.4932	
Mfn2-G12V	27.3779	$P = 0.56078$
Mfn2-G12V+Bax	36.208	
Mfn2(1–703)	58.6826	$P = 0.44327$
Mfn2(1–703)+Bax	61.9097	

Bax/Bak DKO MEFs were transfected with YFP-fused Mfn2 constructs in the presence or absence of Bax and stained with antibodies to cytochrome c. The area of circumscribing Mfn2-YFP per cell was calculated as described in Supplementary Methods. *P* values were calculated by a two-tailed *t*-test analysis of all sample measurements. a.u., arbitrary units.

and Bax/Bak DKO cells transfected with mito-PAGFP were briefly illuminated by a 413-nm laser, resulting in photoactivation of the mitochondrial matrix GFP fluorophore in the ROI. In control 129/CD1 MEFs, dilution and redistribution of mito-PAGFP fluorescence were detected 30 and 60 min after activation (Fig. 5a, c), as previously described in 129/SvEv MEFs¹⁷. Distinct fusion was observed in 42 of 45 cells analysed 30 min after activation and further fusion occurred in 34 of 38 cells analysed 60 min after activation, indicating efficient mitochondrial fusion. By contrast, no dilution or redistribution of mito-PAGFP fluorescence was detectable in 32 of 43 Bax/Bak DKO cells analysed (Fig. 5b, c). Quantifying the extent of fluorescence dilution of GFP after photoactivation in two wild-type MEF lines, Bax/Bak DKO MEFs and Mfn1 KO MEFs, showed that the rate of fusion decreased several fold in the Bax/Bak DKO cells relative to the rate in wild-type MEFs, comparable to the decrease seen in Mfn1 KO cells (Fig. 5c). Similarly, Cos-7 and HeLa cells expressing vMIA showed a complete absence of mitochondrial fusion activity (Supplementary Fig. S6A–C).

In contrast to the defect in mitochondrial morphology, we saw no defect in mitochondrial membrane potential or ATP concentrations in Bax/Bak DKO cells (Supplementary Fig. S14) suggesting a normal metabolic status of Bax/Bak DKO cells. Consistent with this, the amount of hexokinase binding to mitochondria, which seems to be linked to both the rate of glycolysis and adenine nucleotide exchange across the OMM, is unchanged in Bax/Bak DKO MEFs relative to wild-type MEFs¹⁸.

Discussion

Our results show that Bax and Bak are required for normal morpho-

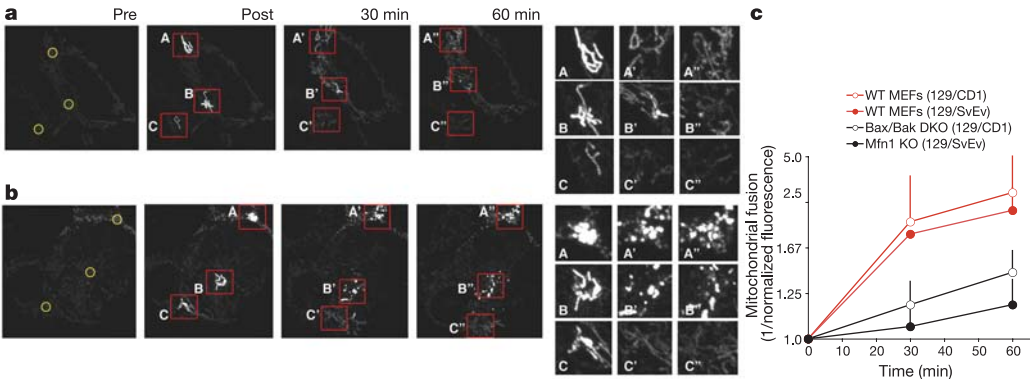


Figure 5 | Mitochondrial fusion is inhibited in the absence of Bax and Bak. **a–c**, Wild-type 129/CD1 (a, c), Bax/Bak DKO (b, c), wild-type 129/SvEv (c) and Mfn1 KO (c) MEFs were analysed for mitochondrial fusion as described¹⁷. Images to the right of **a** and **b** show magnified areas (from red squares) located in close vicinity to the activated ROIs (yellow circles).

c, Quantification of fusion in wild-type 129/CD1, Bax/Bak DKO, wild-type 129/SvEv and Mfn1 KO MEFs as described¹⁷. Data represent the mean $\pm$ s.d. of > 30 single-cell measurements per cell type obtained on three separate occasions.

genesis of mitochondria. The mitochondria in cells lacking Bax and Bak are shorter, have less network continuity and lower fusion rates. Bax also regulates Mfn2 complex assembly, membrane mobility and distribution on the OMM, regulating formation of Mfn2 foci that occur at sites of mitochondrial fusion. In yeast two-hybrid analysis, the 392–602 region of Mfn2, which contains a coiled-coil domain, interacts with Bax (Y.-J. Lee-Wickner and R.Y., unpublished data), suggesting that Bax may directly activate Mfn2. In addition, endophilin B1, a protein normally involved in mitochondrial morphogenesis¹⁴ that binds to Bax²⁰, may coordinate Bax/Bak activity with the mitochondrial morphogenesis machinery.

Our results also yield insights into the role of Bcl-2 family members during apoptosis. On induction of apoptosis, the interconnected mitochondrial network converts into a punctiform morphology²¹. Overexpression of Mfn2 inhibits apoptosis^{15,22}, Mfn2 colocalizes with Bax and Bak in foci during apoptosis^{4,15}, and apoptotic fragmentation of mitochondria occurs in the same time frame as Bax translocation to mitochondria and cytochrome *c* release across the OMM¹⁷, suggesting that mitochondrial structural remodelling participates in transduction of apoptosis. Bax and Bak change conformation during this step of apoptosis, perhaps reflecting inactivation of their mitochondrial elongation activity to allow unbalanced mitochondrial fission; alternatively, because Bax and Bak colocalize with Mfn2 during the fragmentation process, Bax and Bak in their apoptotic conformation may actively inhibit Mfn2 activity to cause Drp1-dependent fragmentation. Similarly, Ced-9, the Bcl-2 family member in *Caenorhabditis elegans*, is thought to convert between anti- and pro-apoptotic conformations and, in concert with Drp1, may regulate both mitochondrial morphology and apoptosis²³. Notably, Ced-9 can induce mitochondrial fusion and co-immunoprecipitates with Mfn2 in mammalian cells²⁴.

METHODS

Measurements of mitochondrial connectivity. FRAP was done as described⁸ to measure mitochondrial connectivity. In brief, bleaching of mitochondrial matrix-targeted YFP fluorescence was applied at randomly chosen perinuclear regions where indicated (bleach), and fluorescence intensity was normalized to the intensity of the ROI from the first image in the series. Differences in the fluorescence recovery between different cell types were measured at 40 s.

Measurements of membrane motility. Circular ROIs, 2.5 µm in diameter, were imaged (α-Plan-FLUAR 100 × /1.45 oil objective; Carl Zeiss) before and after photobleach (20 iterations of 514-nm laser set to 100%) of 0.5-µm circular ROIs located in the centre of imaging area. We collected 150 images (1 scan every 33 ms) and digitalized the fluorescence intensity in imaged ROIs with LSM 510 software (Zeiss MicroImaging). Curves were corrected for both nonspecific photobleaching that occurred during imaging and background, and normalized to the first image in the series (taken as 100%). Each line represents the average of ≥30 measurements obtained on 2–3 separate occasions.

BN-PAGE. Mitochondrial complexes of Mfn2 in transformed wild-type, Bax/Bak DKO and Mfn2 KO MEFs were analysed by BN-PAGE. Mitochondria were isolated by differential centrifugation and incubated at 35 °C for 20 min in the absence or presence of 100 µM GTP-γS (Upstate) and 1 mM GDP (Upstate). After the incubation, samples were centrifuged at 100,000g for 30 min and the resulting pellets were resuspended in membrane lysis buffer (1% CHAPS, 10% glycerol and 0.5 M aminocaproic acid in 50 mM Bis/Tris; pH 7.0), incubated for 20 min on ice, and then centrifuged at 100,000g for 30 min. Supernatants were applied to BN-PAGE at 40 µg of protein per lane. Proteins were resolved on 4–12% gradient polyacrylamide Tris-Glycine gels (Invitrogen) at 4 °C and a current of 3 mA. Proteins were transferred into polyvinylidene fluoride membranes and subjected to immunoblotting with polyclonal antibodies to Mfn2 (a gift from P. Hajek and G. Attardi), monoclonal antibodies to Opa1 (BD Biosciences) and monoclonal antibodies to Hsp60 (Stressgen) as loading controls. The sizes of the complexes in BN-PAGE experiments were established on the basis of the mobility of marker proteins (Amersham): thyroglobulin (669 kDa), apoferritin (440 kDa) and catalase (232 kDa).

Received 7 June; accepted 25 July 2006.

Published online 1 October 2006.

1. Cory, S., Huang, D. C. & Adams, J. M. The Bcl-2 family: roles in cell survival and oncogenesis. *Oncogene* **22**, 8590–8607 (2003).

2. Martinou, J. C. & Green, D. R. Breaking the mitochondrial barrier. *Nature Rev. Mol. Cell Biol.* **2**, 63–67 (2001).
3. Wei, M. C. *et al.* Proapoptotic BAX and BAK: a requisite gateway to mitochondrial dysfunction and death. *Science* **292**, 727–730 (2001).
4. Karbowski, M. *et al.* Spatial and temporal association of Bax with mitochondrial fission sites, Drp1, and Mfn2 during apoptosis. *J. Cell Biol.* **159**, 931–938 (2002).
5. Degenhardt, K., Sundarajan, R., Lindsten, T., Thompson, C. & White, E. Bax and Bak independently promote cytochrome *c* release from mitochondria. *J. Biol. Chem.* **277**, 14127–14134 (2002).
6. Chen, H. *et al.* Mitofusins Mfn1 and Mfn2 coordinately regulate mitochondrial fusion and are essential for embryonic development. *J. Cell Biol.* **160**, 189–200 (2003).
7. Lum, J. J. *et al.* Growth factor regulation of autophagy and cell survival in the absence of apoptosis. *Cell* **120**, 237–248 (2005).
8. Szabadkai, G. *et al.* Drp-1-dependent division of the mitochondrial network blocks intraorganelle Ca²⁺ waves and protects against Ca²⁺-mediated apoptosis. *Mol. Cell* **16**, 59–68 (2004).
9. Arnould, D. *et al.* Cytomegalovirus cell death suppressor vMIA blocks Bax- but not Bak-mediated apoptosis by binding and sequestering Bax at mitochondria. *Proc. Natl Acad. Sci. USA* **101**, 7988–7993 (2004).
10. Poncet, D. *et al.* An anti-apoptotic viral protein that recruits Bax to mitochondria. *J. Biol. Chem.* **279**, 22605–22614 (2004).
11. McCormick, A. L., Smith, V. L., Chow, D. & Mocarski, E. S. Disruption of mitochondrial networks by the human cytomegalovirus UL37 gene product viral mitochondrion-localized inhibitor of apoptosis. *J. Virol.* **77**, 631–641 (2003).
12. van der Bliek, A. M. A mitochondrial division apparatus takes shape. *J. Cell Biol.* **151**, F1–F4 (2000).
13. Shaw, J. M. & Nunnari, J. Mitochondrial dynamics and division in budding yeast. *Trends Cell Biol.* **12**, 178–184 (2002).
14. Karbowski, M., Jeong, S. Y. & Youle, R. J. Endophilin B1 is required for the maintenance of mitochondrial morphology. *J. Cell Biol.* **166**, 1027–1039 (2004).
15. Neuspiel, M., Zunino, R., Gangaraju, S., Rippstein, P. & McBride, H. M. Activated Mfn2 signals mitochondrial fusion, interferes with Bax activation and reduces susceptibility to radical induced depolarization. *J. Biol. Chem.* **280**, 25060–25070 (2005).
16. Koshiba, T. *et al.* Structural basis of mitochondrial tethering by mitofusin complexes. *Science* **305**, 858–862 (2004).
17. Karbowski, M. *et al.* Quantitation of mitochondrial dynamics by photolabeling of individual organelles shows that mitochondrial fusion is blocked during the Bax activation phase of apoptosis. *J. Cell Biol.* **164**, 493–499 (2004).
18. Majewski, N. *et al.* Hexokinase-mitochondria interaction mediated by Akt is required to inhibit apoptosis in the presence or absence of Bax and Bak. *Mol. Cell* **16**, 819–830 (2004).
19. Ishihara, N., Eura, Y. & Mihara, K. Mitofusion 1 and 2 play distinct roles in mitochondria fusion reactions via GTPase activity. *J. Cell Sci.* **117**, 6535–6546 (2004).
20. Takahashi, Y. *et al.* Loss of Bif-1 suppresses Bax/Bak conformational change and mitochondrial apoptosis. *Mol. Cell Biol.* **25**, 9369–9382 (2005).
21. Youle, R. J. & Karbowski, M. Mitochondrial fission in apoptosis. *Nature Rev. Mol. Cell Biol.* **6**, 657–663 (2005).
22. Sugiyama, R., Shimizu, S. & Tsujimoto, Y. Fzo1, a protein involved in mitochondrial fusion, inhibits apoptosis. *J. Biol. Chem.* **279**, 52726–52734 (2004).
23. Jagasia, R., Grote, P., Westermann, B. & Conrad, B. DRP-1-mediated mitochondrial fragmentation during EGL-1-induced cell death in *C. elegans*. *Nature* **433**, 754–760 (2005).
24. Delivani, P., Adrain, C., Taylor, R. C., Duriez, P. J. & Martin, S. J. Role for CED-9 and Egl-1 as regulators of mitochondrial fission and fusion dynamics. *Mol. Cell* **21**, 761–773 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Swarup for help in the early stages of this work; C. Thompson, J. Lum and T. Lindsten for Bax/Bak DKO bone marrow cells, Bak-transfected DKO cells and DKO primary MEFs; E. White, S. Korsmeyer and D. Chan for Bax/Bak DKO BMK cells, Bax/Bak DKO MEFs and Mfn KO MEFs, respectively; B. Vogelstein for HCT116 Bax^{-/-} cells, J. Norris for DU145 cells, V. Goldmacher for vMIA HeLa cells, H. McBride for antibodies to Mfn2 and mutant Mfn2 constructs; P. Hajek and G. Attardi for antibodies to Mfn2; A. Antignani, A. Neutznier, B. Fell, S.-W. Ryu and S. Smith for technical assistance; and C. Smith for assistance with confocal microscopy and data analysis and for reading the manuscript. This work was supported by the Intramural Research Program of the NIH, National Institute of Neurological Disorders and Stroke.

Author Contributions M.K., K.N. and R.Y. contributed to the conception, interpretation, execution and presentation of the experiments; S.-Y.J. contributed to the RNAi experiments; M.C. participated in FRAP experiments and data analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.K. or R.Y. (karbowsk@ninds.nih.gov, youle@ninds.nih.gov).

Structure of eEF3 and the mechanism of transfer RNA release from the E-site

Christian B. F. Andersen^{1*}, Thomas Becker^{2*†}, Michael Blau², Monika Anand³, Mario Halic^{2†}, Bharvi Balar³, Thorsten Mielke⁴, Thomas Boesen¹, Jan Skov Pedersen⁵, Christian M. T. Spahn⁶, Terri Goss Kinzy³, Gregers R. Andersen¹ & Roland Beckmann^{2†}

Elongation factor eEF3 is an ATPase that, in addition to the two canonical factors eEF1A and eEF2, serves an essential function in the translation cycle of fungi. eEF3 is required for the binding of the aminoacyl-tRNA-eEF1A-GTP ternary complex to the ribosomal A-site and has been suggested to facilitate the clearance of deacyl-tRNA from the E-site. Here we present the crystal structure of *Saccharomyces cerevisiae* eEF3, showing that it consists of an amino-terminal HEAT repeat domain, followed by a four-helix bundle and two ABC-type ATPase domains, with a chromodomain inserted in ABC2. Moreover, we present the cryo-electron microscopy structure of the ATP-bound form of eEF3 in complex with the post-translocational-state 80S ribosome from yeast. eEF3 uses an entirely new factor binding site near the ribosomal E-site, with the chromodomain likely to stabilize the ribosomal L1 stalk in an open conformation, thus allowing tRNA release.

Protein synthesis requires, in general, only two canonical GTPase elongation factors. eEF1A (known as EF-Tu in prokaryotes) recruits cognate aminoacyl-tRNAs to the A-site of the ribosome, and, after peptidyl transfer, eEF2 (EF-G in prokaryotes) catalyses translocation of the messenger RNA and the transfer RNAs from the A- and P-sites to the P- and E-sites. In contrast to the canonical factors, eEF3 has an ATPase activity that is stimulated by ribosomes. It interacts with both ribosomal subunits^{1–3}, competes with eEF2 for binding to ribosomes, and stimulates eEF1A-dependent binding of cognate aminoacyl-tRNA to the ribosomal A-site^{1,4}. Because, according to the allosteric three-site model of the ribosomal elongation cycle, E-site release is required for efficient A-site binding, it has been suggested that eEF3 functions as a so-called ‘E-site factor’^{4,5}. Moreover, ATP hydrolysis by eEF3 is required in every elongation cycle to allow chasing of deacyl-tRNA from the E-site⁴.

eEF3 belongs to the family of ABC (ATP-binding cassette) proteins that includes proteins involved in transport across membranes, DNA repair, and translation. The membrane proteins of this class especially represent important targets for development of novel therapeutic strategies. The proteins contain ATP/ADP-binding ABC domains, which convert chemical energy derived from binding of ATP or its hydrolysis into a ‘powerstroke’ of mechanical energy⁶. ABC proteins function as either homodimers or as twin-cassette proteins with two ABC domains within the same polypeptide.

The ribosome exhibits very dynamic behaviour, such as the ratchet movement⁷ or the movement of the L1 and the L7/L12 stalks^{8–11}. Hence, an intriguing question is how the interaction of eEF3 with the ribosome is correlated with its dynamic properties as an ABC protein, and how the energy derived from binding/hydrolysis of ATP is used for its function.

Crystal structure of eEF3

Three crystal structures of residues 1–980 of eEF3 in the apo state (2.7 Å), in complex with ADP (2.4 Å), or in complex with the non-hydrolysable ATP analogue ADPNP (3.0 Å) were solved (Supplementary Table 1) and they have virtually the same overall conformation. The deletion of residues 981–1,044 does not alter the ability of the protein to function as the only form of eEF3¹². The protein has five structural domains (Fig. 1a, b and Supplementary Fig. 1). Residues 1–321 are organized in classical HEAT repeats¹³ with a total of 16 helices. Instead of forming a single long, curved structure as found in PP2A¹⁴, the HEAT domain of eEF3 makes a sharp bend after the first 10 helices. The HEAT domain is connected through a flexible linker to a four-helix bundle (4HB domain; residues 336–413), which packs against the bottom of the concave face of the HEAT domain.

The remaining residues of eEF3 are organized into ABC1 and ABC2 domains homologous to the nucleotide-binding domains of ABC transporters and DNA repair enzymes^{15–18}. These domains each consist of two subdomains (or lobes). Lobe I (residues 418–492 and 563–636 in ABC1, and 637–734 and 915–976 in ABC2; also known as the ATP-binding core) is formed by two β -sheets that wrap around a central α -helix and three exposed helices. Lobe II (493–562 in ABC1, and 734–760 and 870–914 in ABC2; also known as the α -helical subdomain) is composed of four helices in ABC1 and five in ABC2. The arrangement of the two subdomains in both ABC1 and ABC2 is similar to that found in structures of other ABC domains in their apo or ADP-bound states with DALI¹⁹ Z-scores of 13–19, and the Z-score of eEF3 ABC1 versus ABC2 is 15. Hence, the two cassettes in eEF3 are not more similar to each other than to other ABC proteins.

The chromodomain (residues 761–869) is an insert within the α -helical subdomain of ABC2 and contains a five-stranded β -sheet

¹Centre for Structural Biology, Department of Molecular Biology, University of Aarhus, Gustav Wieds Vej 10C, DK-8000 Aarhus, Denmark. ²Institute of Biochemistry, Charité, University Medical School, Humboldt University of Berlin, Monbijoustrasse 2, 10117 Berlin, Germany. ³Department of Molecular Genetics, Microbiology & Immunology, UMDNJ Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5635, USA. ⁴UltraStructureNetwork, USN, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany. ⁵Department of Chemistry and iNANO Interdisciplinary Nanoscience Center, University of Aarhus, Langelandsgade 140, DK-8000 Aarhus, Denmark. ⁶Institut für Medizinische Physik und Biophysik, Charité, University Medical School, Humboldt University of Berlin, Ziegelstrasse 5-9, 10117 Berlin, Germany.

[†]Present address: Gene Center and Department of Chemistry & Biochemistry, University of Munich, Feodor-Lynen-Strasse 25, 81377 Munich, Germany.

*These authors contributed equally to this work.

traversed by an α -helix. Homologous domains from nuclear proteins interact with histones, DNA and RNA²⁰. The ABC1 domain has a limited interface to other domains, whereas the ABC2 domain interacts with the 4HB domain and the N-terminal part of the HEAT domain (Fig. 1a, b). The chromodomain is located at the convex face of the HEAT domain, but makes only a few contacts with it.

In both the ADP and the ADPNP complexes of eEF3, the nucleotide was found 14 Å away from its expected location towards the H-loop of ABC2 (Fig. 1c). Binding to the canonical site is prevented by its overlap with Phe35 from the HEAT domain (Supplementary Fig. 2). In an attempt to disrupt the adenine-binding pocket, we mutated the non-conserved (see Supplementary Fig. 3) Glu 42 to alanine and Ser 82 to tryptophan. These mutations had little effect on yeast growth (data not shown), questioning the physiological relevance of the novel mode of nucleotide binding in

ABC2. No nucleotide is bound in the ABC1 domain, but a single sulphate ion is bound to its P-loop mimicking the α - and β -phosphates. A second sulphate mimics γ -phosphate at the positive helix dipole of the signature motif (Supplementary Figs 1 and 2c), the hallmark of the ABC superfamily. This confirms the previous suggestion that the signature motif binds the inorganic phosphate released by ATP hydrolysis¹⁶.

SAXS analysis of eEF3

Compared with the ATP-bound homodimer structures of MJ0796 and Rad50, the ABC2 domain in eEF3 is rotated approximately 120° away from ABC1. To investigate whether the functionally important ABC tandem formation inferred from other ABC protein structures also takes place in eEF3, we created a model of the ATP state in solution. We maintained the overall location of ABC1 and placed

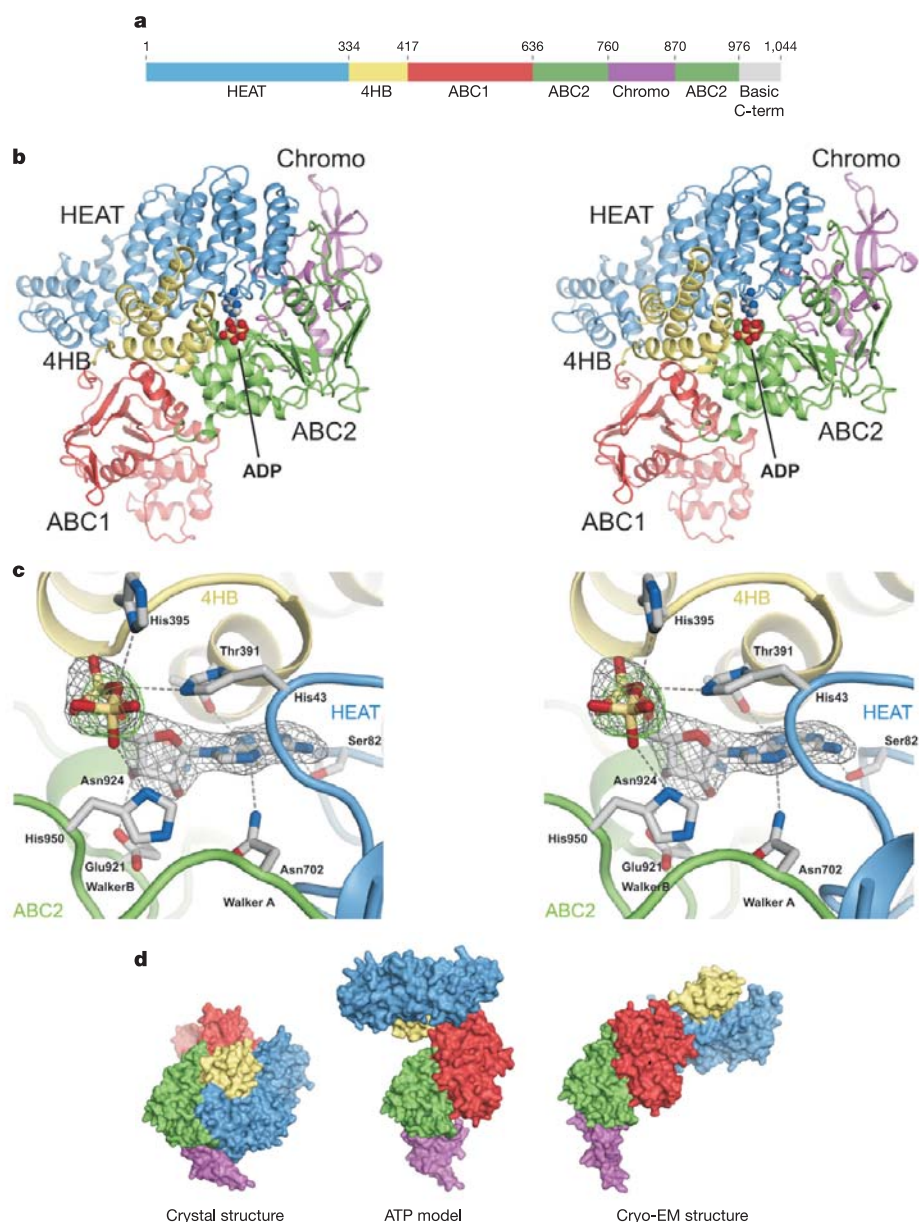


Figure 1 | Structures of *Saccharomyces cerevisiae* eEF3 and nucleotide binding. **a**, Schematic representation of the eEF3 sequence. See the text for a description of the different domains. Chromo, chromodomain; C-term, carboxy-terminal domain. **b**, Stereo view of the crystal structure of eEF3–ADP. The nucleotide is shown in ball-and-stick representation. **c**, Stereo view of the nucleotide-binding site. The electron density of ADP is generated

from an omit map contoured at 1.5 σ (grey) or at 0.8 σ around the β -phosphate (green). **d**, Three different conformations of eEF3. Left: crystal structure of eEF3. Middle: ATP model of eEF3 constructed using homology to the structure of MJ0796. Right: Cryo-electron microscopy (cryo-EM) reconstruction of eEF3 on the 80S ribosome.

ABC2 according to the ATP-induced dimer of MJ0796¹⁶. This was then followed by the rearrangement of the relative location of the two subdomains in both ABC domains (Fig. 1d). This results in an elongated molecule with the chromodomain located at the tip, which, in contrast to the crystal structure, is in good agreement with small-angle X-ray scattering (SAXS) data collected at pH 7.2 (Supplementary Fig. 4). SAXS data collected at the crystallization pH of 5.2 indicate a more compact conformation of eEF3. Hence, in solution a conformational change within eEF3 can be induced by lowering the pH; however, the physiological relevance of this flexibility is unknown.

Reconstitution and cryo-EM of the eEF3-ribosome complex

Purified eEF3 was bound to empty ribosomes or to programmed ribosome nascent chain complexes (RNCs) in the post-translocational (post) state, bearing a peptidyl-tRNA in the P-site²¹. The most stable binding was observed using RNCs in the presence of ADPNP and the antibiotic neomycin¹ (Fig. 2a), whereas empty 80S ribosomes or the presence of ADP (Supplementary Fig. 5) showed weaker binding. As suggested before, this indicates that eEF3 binds to post-state ribosomes, and that ATP hydrolysis is required for eEF3 dissociation from the ribosome⁴.

Cryo-EM and single-particle three-dimensional reconstruction of the RNC-eEF3 complex shows the typical appearance of a programmed yeast ribosome (Fig. 2b) with a P-site tRNA at a resolution of 9.9 Å. Additional density (Supplementary Fig. 6) is present for

eEF3, and consists of two large domains bridging the central protuberance of the 60S and the head of the 40S ribosomal subunit, respectively. In agreement with binding experiments^{2,3,22}, the density of eEF3 could be unambiguously assigned to the HEAT and ABC domains of eEF3, respectively (Fig. 2b).

Ribosomal binding site and conformation of eEF3

The ribosomal binding site for eEF3 is new and completely separated from the binding site for the canonical factors eEF1A and eEF2 (Fig. 2c). The site involves proteins and ribosomal RNA in both the 40S and 60S ribosomal subunit (Fig. 2d; Supplementary Table 2). Non-overlapping binding sites explain how eEF3 and eEF1A can bind simultaneously¹, and confirm that the competition between eEF3 and eEF2 must be a result of different requirements regarding the conformation of the ribosome^{1,4}. Here, eEF3 binds to a post-state ribosome, whereas eEF2 binds to the pre state^{8,23}. The structure of the eEF2-ribosome complex (stabilized by the antibiotic sordarin) indeed reveals a rotated conformation of the head of the 40S subunit^{8,23}, which would hardly allow the binding of eEF3 in the observed conformation.

For building a ribosome-bound eEF3 model (Supplementary Methods), we tried to fit models for eEF3 with the ABC tandem in the ATP state (closed conformation) and ADP state (open conformation). Cross-correlation indicated that eEF3 was indeed present in the ATP state (cross-correlation of 0.78 for ATP versus 0.61 for ADP). The chromodomain had to be moved and rotated relative to its

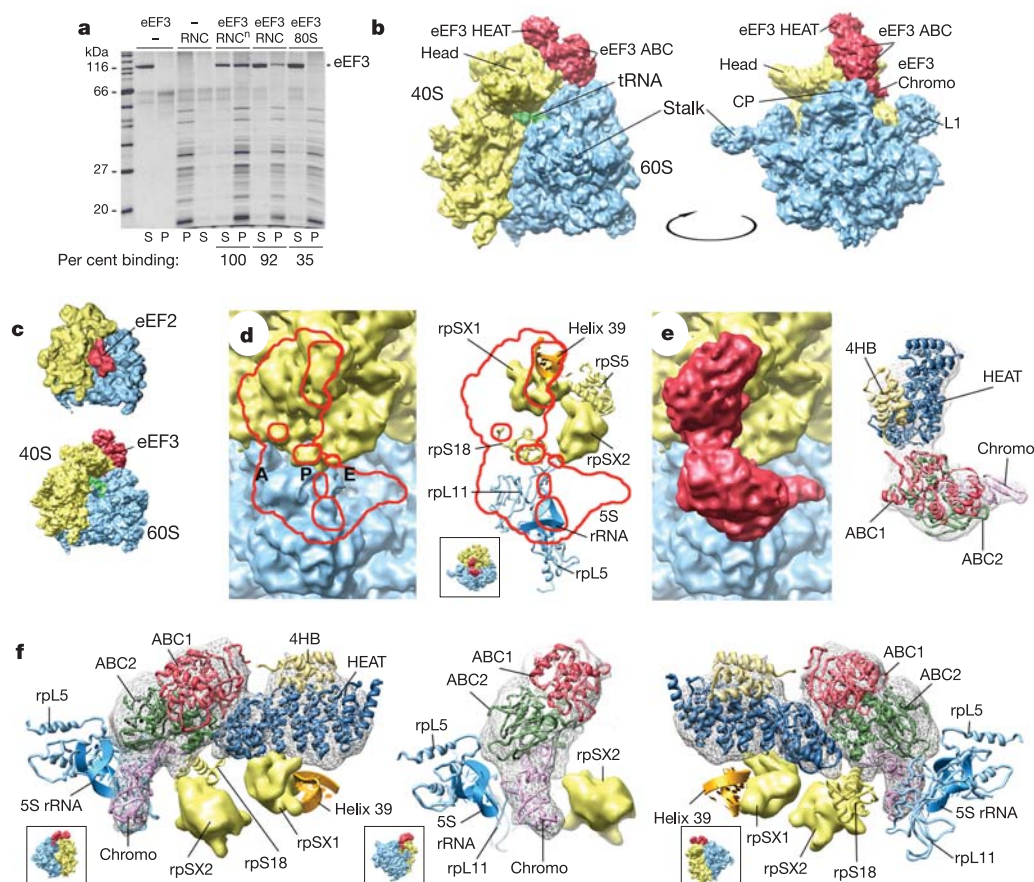


Figure 2 | Reconstitution and cryo-EM structure of the eEF3-ribosome complex. **a**, Binding assay using RNCs or empty 80S ribosomes with eEF3 in the presence of ADPNP, and neomycin (n) where indicated. S, supernatant; P, pellet. **b**, Cryo-EM reconstruction of the eEF3-RNC complex. The right view is rotated by 90° (arrow). CP, central protuberance. **c**, Comparison of the cryo-EM map of the 80S-eEF2-sordarin complex²³ (top) and the eEF3-RNC complex (bottom) reveals different binding sites.

d, Top view (see inset) showing the binding site of eEF3 contoured in red on the ribosomal density (left), and with molecular models of ribosomal components (right). 40S components are shown in yellow, 60S components in blue. Unknown ribosomal proteins (rp) termed rpSX1 and rpSX2 are represented as electron densities. A-, P- and E-sites are indicated by A, P and E. **e**, Density and molecular model of eEF3-ATP on the ribosome. **f**, Molecular contacts of eEF3 with ribosomal components.

position in the crystal structure. The HEAT domain was subdivided into two blocks consisting of 5 and 3 repeats (residues 1–200 and 201–333, respectively), which were docked together with the 4HB domain as individual rigid bodies. Notably, the ribosome-bound conformation of eEF3 (Fig. 1d, right panel) fitted neither the pH 7.2 SAXS scattering curve (Supplementary Fig. 4 and Supplementary Note 1) nor the ATP model (Fig. 1d, middle panel), which suggests that a large conformational change takes place on ribosome binding. This is also observed with other translation factors, such as eEF2²³ and RF2, on ribosome binding. Compared with the ATP-state model, the first five HEAT repeats rotate by 160° in the cryo-EM conformation of eEF3 (Fig. 1d).

The ABC2 domain is positioned on the central protuberance of the large ribosomal subunit, whereas ABC1 does not interact with the ribosome; the chromodomain points towards the ribosomal L1 stalk and is juxtaposed to the E-site (Fig. 2d–f). The ABC2 domain and the chromodomain are anchored to the 60S subunit involving 5S rRNA, rpL11 (known as L5p in *Escherichia coli*) and rpL5 (L18p in *E. coli*) (Supplementary Table 2). Contacts of ABC2 are established between the residues of the first two ABC2 β -strands and a highly conserved region preceding the first β -strand (Supplementary Fig. 1). In the chromodomain, three distinct regions interact with both the 60S and the 40S subunit. Residues 802–808 (representing the Sac7 homology domain, which is known to bind DNA²⁴) contact the 5S rRNA and rpL11 in the 60S subunit. Two linker regions connecting the chromodomain with ABC2 interact with the small ribosomal subunit via rpS18 (S13p in *E. coli*) and an as yet unassigned density (rpSX2) that is likely to correspond to the N-terminal extension of rpS5 (Fig. 2d–f).

The HEAT domain contacts the head of the 40S subunit via the loops connecting the helices of the HEAT repeats. In agreement with binding studies², this interaction involves the tip of helix 39 of the 18S rRNA, an unknown ribosomal protein (here termed rpSX1) and rpS18 (Supplementary Table 2; Fig. 2f). HEAT repeats active in RNA binding have only been observed for the Ro autoantigen, where the helix-connecting loops contact single-stranded RNA via basic amino-acid residues²⁵. Likewise, all of the contacts formed between the HEAT domain of eEF3 and ribosomal RNA (and also the ribosomal proteins) involve loops.

Involvement of eEF3 in tRNA release

The question remains of how eEF3 acts on the ribosome as a dynamic ABC protein and whether this action can be related to its suggested function as an E-site factor. It is intriguing that the chromodomain of eEF3 is positioned in the immediate proximity of the ribosomal E-site where it is capable of influencing the head of the 40S subunit and the L1 stalk, both of which contribute to the affinity of tRNA for the E-site^{23,26,27}. It is very likely that a conformational switch of the chromodomain occurs on ribosome binding. Such a switch would correlate well with the movement of the L1 stalk from an E-site-secluding 'in' position to an 'open' position, which would, thus, unlock the E-site to allow release of the tRNA (Fig. 3). An important function of the chromodomain is further supported by its placement between the Q-loop and the signature motif. In BtuCD the corresponding region interacts with the transmembrane domain¹⁵, and in Rad50¹⁸ this region harbours the coiled-coil forming the ring around two strands of DNA, both of which undergo conformational changes. Likewise, ATP binding and hydrolysis by eEF3 might induce a change in the relative orientation of the two subdomains in ABC2, and thereby change the position of the chromodomain. We propose that this could, in turn, influence the conformation of the L1 stalk.

Conclusions

Our findings demonstrate conformational flexibility of eEF3 that is likely to be used for its function. In the eEF3–ADPNP conformation bound to the 80S ribosome (Fig. 1d), the two ABC domains dimerize, with the ABC tandem being present in the ATP-bound closed dimer

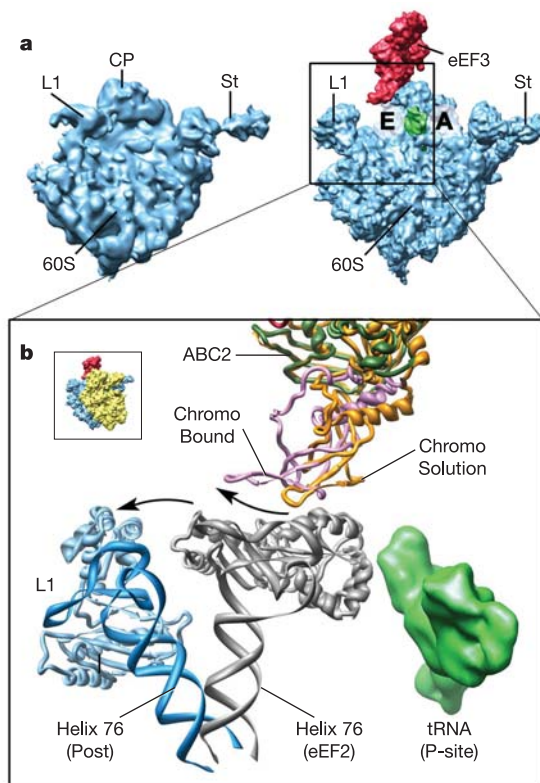


Figure 3 | Dynamic behaviour of eEF3 and the 80S ribosome. **a**, Position of eEF3 with respect to the ribosomal E- and A-sites (indicated by E and A), and to the dynamic L1 stalk (St) in the 'in' and 'out' position as observed in the 80S–eEF2–sordarin complex (left) and the eEF3–RNC complex (right), respectively. **b**, Movement of the chromodomain of eEF3 and the L1 stalk of the 60S subunit. Arrows indicate the movements of the L1 stalk from the eEF2 state (grey) to the post state (blue), and the switch of the chromodomain from a solution-state model (yellow) to the observed ribosome-bound (ATP) state (magenta).

state observed in the ATP-bound structures of Rad50 and MJ0796. Compared with our solution-state model, a large reorientation of the HEAT and 4HB domains relative to ABC1 has occurred (Fig. 1d). So far, structures of homodimeric ABC proteins, except for the low-resolution structures of MsaA^{17,28}, have indicated only relatively small differences between the open and closed dimeric states. The crystal structure of eEF3 demonstrates that the nucleotide-binding domains of twin-cassette ABC proteins may undergo very large intramolecular conformational changes. In eEF3, this could be required especially during binding and release from the ribosome. The SAXS data collected at pH 7.2 showed that the ABC1–ABC2 domain interface most likely pre-forms in solution. The weak intrinsic ribosome-independent ATPase activity²⁹ suggests that the ABC tandem is somewhat open in solution. This is then likely to rearrange to a tighter tandem with higher ATPase activity on ribosome binding, in agreement with the 36-fold stimulation of eEF3 ATPase activity by the ribosome²⁹.

We suggest that the chromodomain stabilizes the L1 stalk in the 'out' position (Fig. 3b). Rearrangement of L1 is a prerequisite for the release of the tRNA from the E-site, as this site is secluded with L1 in the 'in' position^{10,23}. Therefore, one aspect of eEF3 function may be the opening of the E-site by moving the L1 stalk. Notably, the largest movements of the L1 stalk have been observed in *S. cerevisiae*²³, an organism—as are all other fungi—strictly dependent on eEF3 activity. An essential function of the chromodomain on the ribosome explains why proteolytic degradation of this domain preserves the intrinsic ATPase activity but destroys the ribosome-dependent stimulation²².

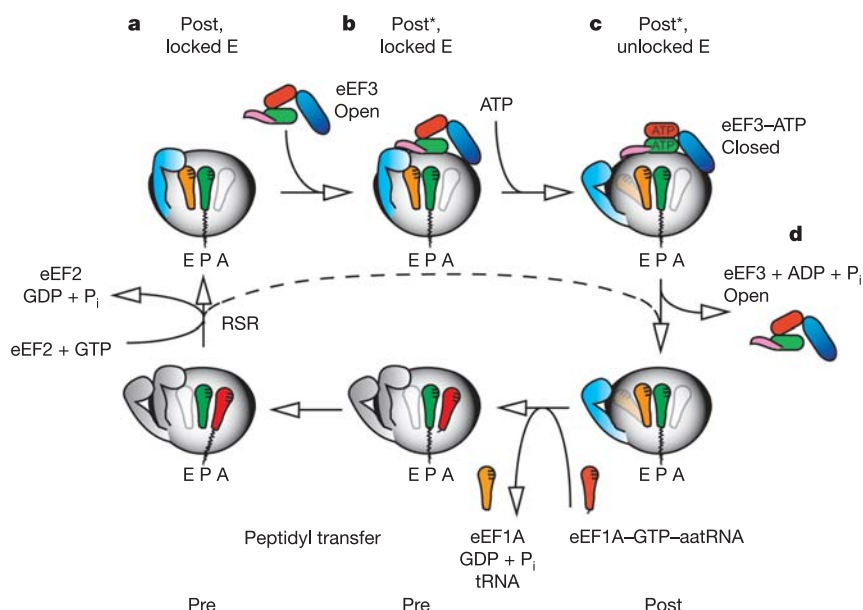


Figure 4 | Model of the role of eEF3 in the fungal elongation cycle. **a**, The post-state ribosome with a locked E-site tRNA owing to the L1 stalk in the 'in' position and the conformation of the 40S head (Post, locked E). **b**, Hypothetical initial interaction of eEF3 in the open tandem or intermediate conformation (Post*, locked E). **c**, Ribosome interaction triggers the ATP-dependent closed tandem formation and high-affinity ribosome binding by eEF3, as observed by cryo-EM (Post*). **d**, ATP hydrolysis of the closed tandem results in the dissociation of eEF3, E-site opening, and unlocking of the 40S head (Post). Now, eEF1A-GTP-aminoacyl-tRNA can bind and the E-site deacyl-tRNA is released. ATP hydrolysis by eEF3, tRNA release, and A-site loading by eEF1A may take place as a joint event. aa-tRNA, aminoacyl-tRNA. RSR, ratchet-like subunit rearrangement.

The conformational switch of eEF3 could also lead to a change of the 40S head conformation via the HEAT domain, resulting in altered properties of the E-site^{23,26,27}. This would also follow the general principle of ABC transporters, where the rearrangement of the two ABC domains between an open and a closed dimer state is used for the rearrangement of the domains connected to them. eEF3 itself is partially closing the E-site and stalls the head of the 40S subunit in its position. Hence, it is likely that tRNA release is dependent on the dissociation of eEF3 and release of the stalled conformation of the 40S subunit head. This would explain why efficient chasing from the E-site with cognate deacyl-tRNA is dependent on ATP hydrolysis by eEF3⁴, which would trigger the dissociation of eEF3 from the ribosome.

Following the switch model for ABC transporters, in which substrate binding triggers ATP binding³⁰, eEF3 binds to the post-state ribosomes also in the presence of ADP (Supplementary Fig. 5), indicative of the open conformation of the ABC tandem. On this basis, we suggest a model describing the role of eEF3 in the fungal translation cycle, which is expanded by fungi-specific post states of the ribosome (Fig. 4). (1) After eEF2-induced tRNA translocation, the conformation of the 40S head and of the L1 stalk in the 'in' position maintains the E-site tRNA as secluded (Fig. 4, "Post, locked E"). (2) Initial weak interaction of eEF3 with the ribosome may occur in the open ABC tandem, or an intermediate conformation of the ABC domains, without inducing changes in the ribosome ("Post*, locked E"). (3) The interaction with the ribosome may trigger the ATP-dependent closed tandem formation and high-affinity ribosome binding by eEF3 in the conformation observed by cryo-EM. A conformational switch of the eEF3 chromodomain may stabilize the L1 stalk in the 'out' position, thereby opening the E-site ("Post*, unlocked E"). Now, eEF3 itself partially closes the E-site and fixes the 40S head. (4) The closed tandem elicits ATP hydrolysis, causing the dissociation of eEF3 from the ribosome. This unlocks the 40S head and results in an open E-site from which the tRNA can be released, probably owing to an increased off-rate ("Post"). Now, aminoacyl-tRNA-eEF1A-GTP can deliver the next aminoacyl-tRNA to the

A-site. According to the allosteric three-site model⁴, a closed E-site preventing release of deacyl-tRNA would interfere with A-site binding. As eEF3 and eEF1A can bind simultaneously to the ribosome, ATP hydrolysis by eEF3, tRNA release, and A-site loading by eEF1A may take place as a joint event.

Taken together, we have provided the structural basis for understanding the essential function of eEF3, which will be instrumental for the future dissection of the exact molecular mechanism of this unique fungal factor, and, in addition, may be used for rational drug design.

METHODS

An initial polyaniline model was built into the 3.5 Å resolution density map of eEF3-ADP¹². This model was used for molecular replacement with the 2.4 Å selenomethionine data set. Phase combination allowed the construction of a new polyaniline model by RESOLVE³¹, rebuilding in O³², and refinement with CNS³³ (Supplementary Table 1). Crystals of apo eEF3 and eEF3-ADPNP were obtained using the same approach as eEF3-ADP¹² by co-crystallization in the presence of 45% (w/v) ammonium sulphate at pH 5.2, and data processed with XDS³⁴ (Supplementary Table 1). The eEF3-ADP structure was used for rigid-body refinement against these data, and the resulting model was manually refitted in O and refined in CNS (Supplementary Table 1).

The SAXS measurements were performed³⁵ at pH 5.2 or pH 7.2, and 1 mM ADP/ATP when present. Simulated SAXS scattering curves were calculated with CRYSOLE³⁶, and experimental molecular envelopes with GASBOR³⁷.

The 80S ribosomes and RNCs were purified as described before^{21,38,39}. eEF3-Xa-His₁₀ was overexpressed in yeast and affinity-purified. 80S ribosome-eEF3 complexes were reconstituted and applied on carbon-coated holey grids⁴⁰ in the presence of Sec61²¹, visualized on a Polara cryo-microscope, and processed using SPIDER⁴¹. The data set was sorted according to empty³⁸ or programmed²¹ conformational state, followed by sorting according to the presence of eEF3. A total of 37,700 particles were used for the final reconstruction at a resolution of 9.9 Å (6.2 Å) based on Fourier shell correlation of masked volumes at 0.5 (3σ). The map is shown without the Sec61 density in Figs 2 and 3 for better clarity (see Supplementary Fig. 6a for the complete map).

Docking of X-ray structures and molecular models of eEF3^{42,43} and ribosomal proteins/RNA⁴⁴ was done with IRIS Explorer, SPIDER⁴¹, O³² and ERNA-3D⁴⁵. Using the program Situs⁴⁶ for docking of MalK-based⁴⁷ ATP and ADP models for

eEF3 (cross-correlation of 0.7 and 0.62, respectively) confirmed our results with models based on MJ0796⁴² and RLI⁴³. Figures were produced with PYMOL⁴⁸ and USFC Chimera⁴⁹.

Received 10 October 2005; accepted 3 August 2006.

Published online 23 August 2006.

- Kovalchuk, O., Kambampati, R., Pladies, E. & Chakraborty, K. Competition and cooperation among yeast elongation factors. *Eur. J. Biochem.* **258**, 986–993 (1998).
- Gontarek, R. R., Li, H., Nurse, K. & Prescott, C. D. The N terminus of eukaryotic translation elongation factor 3 interacts with 18S rRNA and 80S ribosomes. *J. Biol. Chem.* **273**, 10249–10252 (1998).
- Kambampati, R. *et al.* Limited proteolysis of yeast elongation factor 3. Sequence and location of the subdomains. *J. Biol. Chem.* **275**, 16963–16968 (2000).
- Triana-Alonso, F. J., Chakraborty, K. & Nierhaus, K. H. The elongation factor 3 unique in higher fungi and essential for protein biosynthesis is an E site factor. *J. Biol. Chem.* **270**, 20473–20478 (1995).
- Kamath, A. & Chakraborty, K. Role of yeast elongation factor 3 in the elongation cycle. *J. Biol. Chem.* **264**, 15423–15428 (1989).
- Jones, P. M. & George, A. M. Mechanism of ABC transporters: a molecular dynamics simulation of a well characterized nucleotide-binding subunit. *Proc. Natl Acad. Sci. USA* **99**, 12639–12644 (2002).
- Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406**, 318–322 (2000).
- Gomez-Lorenzo, M. G. *et al.* Three-dimensional cryo-electron microscopy localization of EF2 in the *Saccharomyces cerevisiae* 80S ribosome at 17.5 Å resolution. *EMBO J.* **19**, 2710–2718 (2000).
- Valle, M. *et al.* Locking and unlocking of ribosomal motions. *Cell* **114**, 123–134 (2003).
- Harms, J. *et al.* High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107**, 679–688 (2001).
- Morgan, D. G., Menetret, J. F., Neuhaus, A., Rapoport, T. A. & Akey, C. W. Structure of the mammalian ribosome–channel complex at 17 Å resolution. *J. Mol. Biol.* **324**, 871–886 (2002).
- Andersen, C. F. *et al.* Purification and crystallization of the yeast translation elongation factor eEF3. *Acta Crystallogr. D* **60**, 1304–1307 (2004).
- Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W. & Bork, P. Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309**, 1–18 (2001).
- Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A. & Barford, D. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* **96**, 99–110 (1999).
- Locher, K. P., Lee, A. T. & Rees, D. C. The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* **296**, 1091–1098 (2002).
- Smith, P. C. *et al.* ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10**, 139–149 (2002).
- Chang, G. & Roth, C. B. Structure of MsbA from *E. coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **293**, 1793–1800 (2001).
- Hopfner, K. P. *et al.* Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell* **101**, 789–800 (2000).
- Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–603 (1996).
- Brehm, A., Tufteland, K. R., Aasland, R. & Becker, P. B. The many colours of chromodomains. *Bioessays* **26**, 133–140 (2004).
- Beckmann, R. *et al.* Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372 (2001).
- Kambampati, R. & Chakraborty, K. Functional subdomains of yeast elongation factor 3. Localization of ribosome-binding domain. *J. Biol. Chem.* **272**, 6377–6381 (1997).
- Spahn, C. M. *et al.* Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. *EMBO J.* **23**, 1008–1019 (2004).
- Robinson, H. *et al.* The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature* **392**, 202–205 (1998).
- Stein, A. J., Fuchs, G., Fu, C., Wolin, S. L. & Reinisch, K. M. Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. *Cell* **121**, 529–539 (2005).
- Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
- Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
- Reyes, C. L. & Chang, G. Structure of the ABC transporter MsbA in complex with ADP·vanadate and lipopolysaccharide. *Science* **308**, 1028–1031 (2005).
- Sarthy, A. V. *et al.* High-level overexpression of yeast elongation factor 3 and detailed kinetic analysis using a coupled spectrophotometric assay. *Anal. Biochem.* **254**, 288–290 (1997).
- Higgins, C. F. & Linton, K. J. The ATP switch model for ABC transporters. *Nature Struct. Mol. Biol.* **11**, 918–926 (2004).
- Terwilliger, T. SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchrotron Radiat.* **11**, 49–52 (2004).
- Jones, T. A., Cowan, S., Zou, J.-Y. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Kabsch, W. in *International Tables for Crystallography* (ed. Arnold, E.) Ch. 25.2.9 (Kluwer Academic, Dordrecht, 2001).
- Pedersen, J. S. A flux- and background-optimized version of the NanoSTAR small-angle X-ray scattering camera for solution scattering. *J. Appl. Cryst.* **37**, 369–380 (2004).
- Svergun, D. I., Barberato, C. & Koch, M. H. J. CRY SOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* **28**, 768–773 (1995).
- Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953 (2001).
- Beckmann, R. *et al.* Alignment of conduits for the nascent polypeptide chain in the ribosome–Sec61 complex. *Science* **278**, 2123–2126 (1997).
- Halic, M. *et al.* Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* **427**, 808–814 (2004).
- Wagenknecht, T., Grassucci, R. & Frank, J. Electron microscopy and computer image averaging of ice-embedded large ribosomal subunits from *Escherichia coli*. *J. Mol. Biol.* **199**, 137–147 (1988).
- Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
- Yuan, Y. R. *et al.* The crystal structure of the MJ0796 ATP-binding cassette. Implications for the structural consequences of ATP hydrolysis in the active site of an ABC transporter. *J. Biol. Chem.* **276**, 32313–32321 (2001).
- Karcher, A., Buttner, K., Martens, B., Jansen, R. P. & Hopfner, K. P. X-ray structure of RLI, an essential twin cassette ABC ATPase involved in ribosome biogenesis and HIV capsid assembly. *Structure (Camb.)* **13**, 649–659 (2005).
- Spahn, C. M. *et al.* Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA–ribosome and subunit–subunit interactions. *Cell* **107**, 373–386 (2001).
- Mueller, F. *et al.* Getting closer to an understanding of the three-dimensional structure of ribosomal RNA. *Biochem. Cell Biol.* **73**, 767–773 (1995).
- Wriggers, W., Milligan, R. A. & McCammon, J. A. Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125**, 185–195 (1999).
- Lu, G., Westbrook, J. M., Davidson, A. L. & Chen, J. ATP hydrolysis is required to reset the ATP-binding cassette dimer into the resting-state conformation. *Proc. Natl Acad. Sci. USA* **102**, 17969–17974 (2005).
- DeLano, W. L. *The PyMOL User's Manual* (DeLano Scientific, San Carlos, California, 2002).
- Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements G.R.A. was supported by the NIH, HFSP, EU FP5 and SNF. T.G.K. was supported by the NIH and HFSP. We are grateful to I. Kerr and D. Brodersen for discussions, and to L. Van for technical assistance. The work was further supported by grants from the VolkswagenStiftung (to R.B. and C.M.T.S.), the Deutsche Forschungsgemeinschaft (to R.B.), and by the European Union and Senatsverwaltung für Wissenschaft, Forschung und Kultur Berlin (UltraStructureNetwork).

Author Information Coordinates and structure factors for the crystal structures have been deposited in the RCSB Protein Data Bank under accession numbers 2IX3, 2IW3 and 2IWH; the electron microscopy (EM) map has been deposited in the EMBL-European Bioinformatics Institute EM Data Bank under accession number EMD-1233 and coordinates of EM-based models in the RCSB Protein Data Bank under accession number 2IX8. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.B. (beckmann@lmb.uni-muenchen.de) and G.R.A. (gra@mb.au.dk).

The sequestration of ethane on Titan in smog particles

D. M. Hunten¹

Saturn's largest satellite, Titan, has a dense atmosphere of nitrogen with a few per cent of methane¹. At visible wavelengths its surface is hidden by dense orange-brown smog, which is produced in the stratosphere by photochemical reactions following the dissociation of methane by solar ultraviolet light. The most abundant of the products of these reactions is ethane, and enough of it should have been generated over the life of the Solar System to form a satellite-wide ocean one kilometre deep². Radar observations³ have found specular reflections in 75 per cent of the surface spots observed, but optical searches for a sun-glint off an ocean have been negative⁴. Here I explain the mysterious absence or rarity of liquid ethane: it condenses onto the smog particles, instead of into liquid drops, at the cold temperatures in Titan's atmosphere. This dusty combination of smog and ethane, forming deposits several kilometres thick on the surface, including the observed dunes and dark areas, could be named 'smust'. This satellite-wide deposit replaces the ocean long thought to be an important feature of Titan.

Smog particles and ethane molecules exist in Jupiter's atmosphere as well as Titan's, and their behaviour on Jupiter gives important clues for the understanding of Titan. On Jupiter they are less abundant than on Titan because of the large quantity of hydrogen present, which is responsible for some recycling of methane from its dissociation products⁵, but this small difference does not affect the important point: the vertical distribution of ethane in Jupiter's atmosphere requires that it should condense on the hydrocarbon smog particles. It is released only at temperatures above 300 K or, in the stratosphere, by photon-stimulated desorption. At the very low temperatures of Titan's lower atmosphere, the ethane (as well as other non-methane hydrocarbons) remains bound to the smog particles and is therefore not available for condensation into liquid ethane.

Ethane in Jupiter's troposphere was measured by the Galileo Probe Mass Spectrometer^{6,7}, and its mixing ratio at 10–15 bar (6×10^{-6}) is much greater than it is near 1 bar (4×10^{-8}). A source is required in the deeper (higher-pressure) region, and it was suggested⁷ to be evaporation and dissociation, at the relatively warm temperatures of that level, of molecules from the smog particles. Here I note that the ethane mixing ratio of 10^{-5} measured in the stratosphere (at 10^{-3} bar) by its infrared emission⁸ is also much greater than at 1 bar, and that the small value of 4×10^{-8} near the latter level represents a deep minimum. The temperature is cold, about 150 K, but not cold enough to liquefy ethane. The minimum can be explained by the condensation of ethane molecules on (and in) the smog particles. They are released at and below Jupiter's 10 bar level, along with the original smog constituents, by the higher temperatures there: 350 K and hotter at deeper levels. Stratospheric temperatures, although warmer than those near 1 bar, are still not warm enough to release the observed ethane molecules; instead they are

undoubtedly released by ultraviolet photons which do not penetrate to the 1 bar region on Jupiter, and the same must be true on Titan.

All these processes (except evaporation at deep levels) should operate in Titan's atmosphere with its much lower temperatures. Condensation of ethane into a liquid requires a sufficiently high concentration, which cannot be attained because ethane is more strongly bound to the smog particles. The lack of an ocean or other large bodies of liquid thus has a natural explanation.

The name I suggest for the material that accumulates on the surface is 'smust', a contraction of 'smog + dust'. The basic point of this paper depends only on the observations of the vertical distribution of ethane in Jupiter's atmosphere and not on any other properties, which are discussed below on the basis of observations in Titan's atmosphere from the Huygens probe and on photochemical theory confirmed by remote observations of hydrocarbon molecules in the atmosphere.

The particles, observed in detail by the DISR (Descent Imager and Spectral Radiometer) instrument⁹ on Huygens, were found to have a radius of around $0.9 \mu\text{m}$. Even if they form larger clumps once they have been deposited on the surface, they are more like 'dust' than 'sand'. The amount generated over the life of the Solar System can be obtained from published theoretical calculations, the same ones that were used² to find an amount for the ethane in the proposed ocean. Depending on the assumed density, the depth of a uniform layer could be several kilometres. The optical properties of a particle as observed from the Huygens probe⁹ are attributed to a loose aggregate of monomers with a radius of $0.05 \mu\text{m}$. Each particle contains 256 or 512 monomers; the following description adopts the 512 value.

The radius of a sphere having the same surface area as an aggregate particle is $0.9 \mu\text{m}$, but the loose, fractal structure of the particles implies a considerably greater physical dimension. For an estimate of the composition it is necessary to turn to photochemical computations⁵ of production rates. The basic composition is assumed to be of polyyne (polyacetylene) molecules ranging from C_2H_2 to C_8H_2 ; the production rate of longer carbon chains is believed to be sufficiently smaller that such chains can be neglected. Acetylene (C_2H_2) is assumed here to behave like the ethane. Following the evidence discussed above, ethane is included in the mix; it accounts for 48% of the mass while the four polynes amount to 10, 14, 26 and 2%, respectively. (Ethylene, C_2H_4 , is observed and computed to be much rarer, because it is readily photolysed.) The mean mass is 45.4 atomic mass units. The volume of an individual monomer, assumed spherical, is $5.24 \times 10^{-16} \text{cm}^3$; if the density is assumed to be 0.1g cm^{-3} the mass is $5.2 \times 10^{-17} \text{g}$ and the number of molecules is 7×10^5 . The corresponding numbers for a particle of 512 monomers are $3 \times 10^{-14} \text{g}$ and 4×10^8 molecules.

The total flux down to the surface may be estimated from the values for the individual molecules⁵; starting with ethane, they are (in

¹Lunar and Planetary Laboratory, The University of Arizona, Tucson, Arizona 85721, USA.

molecules $\text{cm}^{-2}\text{s}^{-1}$): 5.8×10^9 , 1.2×10^9 , 1.7×10^9 , 3.1×10^9 and 2.5×10^8 . Weighting each of these by the number of carbon atoms in the molecule gives a total flux of 4.1×10^{10} carbon atoms $\text{cm}^{-2}\text{s}^{-1}$; if this flux is assumed to have been constant for the life of the Solar System, the accumulated quantity is 5.9×10^{27} carbon atoms cm^{-2} and the surface density of carbon is $1.2 \times 10^5 \text{ g cm}^{-2}$; adding an allowance of one hydrogen atom per carbon atom increases the surface density of the deposit to $1.3 \times 10^5 \text{ g cm}^{-2}$. Undoubtedly most of the material is compressed by the mass lying above it; if the density is assumed to be 0.5 g cm^{-2} the average depth of the deposit is 2.6 km and its total volume $2 \times 10^8 \text{ km}^3$.

Widespread dunes have been observed by the Cassini radar. The paper reporting them¹⁰ proposes, as one option, that the origin of dunes lies in "Titan's atmospheric methane photochemistry" and gives an estimate of the volume of the material smaller than the one above by a factor of 20. Here I make the similar proposal that smust is the material of the dunes, and in addition that it contains the missing ethane and thus accounts for the rarity of liquid surfaces. The grain diameter for formation of dunes was suggested¹⁰ to be 180–250 μm , far larger than the 1.8 μm inferred for the grains in the atmosphere from the Huygens observations. However, if the structure of these aggregates is very loose the outside dimension could be considerably greater. The compressive forces in a deep layer may cause these original smust particles to aggregate into larger ones, which would naturally migrate to the surface as the small ones fill the voids in the material (the so-called 'Brazil nut effect'). Only a small fraction of the particles in the smust deposit need be larger than those observed by Huygens, and would still be plenty to account for the dunes.

Alternatively, the particles in the dunes may be much smaller than the 180–250 μm , which is the optimum¹¹ size for dune formation. Smaller particles can still be raised by the wind if the speeds are greater, especially if the interparticle cohesive forces are smaller than was assumed—it is these forces that inhibit the smallest particles from being lofted by the wind. The particles near the surface are found⁹ to be rather different from those at higher altitudes; they may include a component raised from the surface in addition to those falling from the stratosphere.

It would be desirable to verify in the laboratory the attachment of ethane molecules to smog particles, here deduced from their behaviour on Jupiter. Such particles, with their fluffy structure, have, however, not been produced in experiments, which instead generate a dense deposit on the walls of the vessel. It will be a challenge to reproduce this structure along with a realistic composition, and then to expose the particles to ethane molecules.

Note added in proof: A few days before this paper went to press, a report¹² appeared suggesting the presence of a tenuous ethane cloud bordering the northern (winter) polar region and perhaps extending over the rest of the polar cap, which is in darkness. It occupies the height region of 30–50 km and contains a column of particles of

radius 3 μm containing 60,000 particles cm^{-2} : one particle per 33 cm in the 1 cm^2 column. The evidence that the particles are composed of ethane is indirect but convincing. Their presence may be compatible with the smust particles discussed here, because they contain only a tiny fraction of the total ethane. The mixing ratio quoted¹² for the stratosphere is 2.2×10^{-5} . The number of ethane molecules in the proposed cloud is $7.6 \times 10^{16} \text{ cm}^{-2}$; the corresponding number of molecules in the ambient atmospheric column from 30 to 50 km is $4.1 \times 10^{25} \text{ cm}^{-2}$ and the ethane mixing ratio is therefore 1.8×10^{-9} , only 8×10^{-5} of the available 2.2×10^{-5} . It is entirely reasonable that these few molecules would not reside on the smust particles. A possible difficulty is that this small amount of ethane vapour would be unable to condense. On the other hand, if more ethane were available one would expect the cloud particles to grow larger; probably the attachment of most of the ethane to the smust particles is necessary to prevent this.

Received 25 June; accepted 1 August 2006.

1. Niemann, H. B. *et al.* The abundances and constituents of Titan's atmosphere from the GCMS instrument on the Huygens probe. *Nature* **438**, 779–784 (2005).
2. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
3. Campbell, D. B., Black, G. J., Carter, L. M. & Ostro, S. J. Radar evidence for liquid surfaces on Titan. *Science* **302**, 431–434 (2003).
4. West, R. A., Brown, M. E., Salinas, S. V., Bouchez, A. H. & Roe, H. G. No oceans on Titan from the absence of a near-infrared specular reflection. *Nature* **436**, 670–672 (2005).
5. Yung, Y. L. & DeMore, W. B. *Photochemistry of Planetary Atmospheres* Ch. 5, 6.3 (Oxford Univ. Press, New York, 1999).
6. Niemann, H. B. *et al.* The Galileo Probe Mass Spectrometer: composition of Jupiter's atmosphere. *Science* **272**, 846–849 (1996).
7. Niemann, H. B. *et al.* The composition of the Jovian atmosphere as determined by the Galileo Probe Mass Spectrometer. *J. Geophys. Res.* **103**, 22831–22846 (1998).
8. Fouchet, T. *et al.* Jupiter's hydrocarbons observed with ISO-SWS: vertical profiles of C_2H_6 and C_2H_2 , detection of $\text{CH}_3\text{C}_2\text{H}$. *Astron. Astrophys.* **355**, L13–L17 (2000).
9. Tomasko, M. G. *et al.* Rain, winds and haze during the Huygens probe's descent to Titan's surface. *Nature* **438**, 765–778 (2005).
10. Lorenz, R. D. *et al.* The sand seas of Titan: Cassini RADAR observations of longitudinal dunes. *Science* **312**, 724–727 (2006).
11. Lorenz, R. D., Lunine, J. I., Grier, J. A. & Fisher, M. A. Prediction of aeolian features on planets: application to Titan. *J. Geophys. Res.* **100**, 26377–26386 (1995).
12. Griffith, C. A. *et al.* Evidence for a polar ethane cloud on Titan. *Science* **313**, 1620–1622 (2006).

Acknowledgements I thank A. L. Sprague, L. Doose and M. Tomasko for discussions. I also thank H. Niemann, the Principal Investigator of the Galileo Probe Mass Spectrometer.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to the author (dhunten@lpl.arizona.edu).

Observation of strong coupling between one atom and a monolithic microresonator

Takao Aoki^{1†}, Barak Dayan¹, E. Wilcut¹, W. P. Bowen^{1†}, A. S. Parkins^{1†}, T. J. Kippenberg^{2†}, K. J. Vahala² & H. J. Kimble¹

Over the past decade, strong interactions of light and matter at the single-photon level have enabled a wide set of scientific advances in quantum optics and quantum information science. This work has been performed principally within the setting of cavity quantum electrodynamics^{1–4} with diverse physical systems⁵, including single atoms in Fabry–Perot resonators^{1,6}, quantum dots coupled to micropillars and photonic bandgap cavities^{7,8} and Cooper pairs interacting with superconducting resonators^{9,10}. Experiments with single, localized atoms have been at the forefront of these advances^{11–15} with the use of optical resonators in high-finesse Fabry–Perot configurations¹⁶. As a result of the extreme technical challenges involved in further improving the multilayer dielectric mirror coatings¹⁷ of these resonators and in scaling to large numbers of devices, there has been increased interest in the development of alternative microcavity systems⁵. Here we show strong coupling between individual caesium atoms and the fields of a high-quality toroidal microresonator. From observations of transit events for single atoms falling through the resonator's evanescent field, we determine the coherent coupling rate for interactions near the surface of the resonator. We develop a theoretical model to quantify our observations, demonstrating that strong coupling is achieved, with the rate of coherent coupling exceeding the dissipative rates of the atom and the cavity. Our work opens the way for investigations of optical processes with single atoms and photons in lithographically fabricated microresonators. Applications include the implementation of quantum networks^{18,19}, scalable quantum logic with photons²⁰, and quantum information processing on atom chips²¹.

The realization of large-scale quantum networks^{18,19} requires the capability to inter-connect many 'quantum nodes', each of which could consist of a microresonator containing a set of trapped atoms. The 'quantum channels' to connect these nodes would be optical fibres, with strong interactions in cavity quantum electrodynamics (QED) providing an efficient interface between light and matter. Here we provide a critical step towards a feasible quantum network by demonstrating strong coupling of single atoms to microresonators fabricated on silicon wafers in large numbers by standard lithographic techniques followed by a laser-reflow process²². Combined with the capability to couple light efficiently to and from such cavities directly via a tapered optical fibre²³, toroidal microcavities offer promising capabilities for new nonlinear interactions of single atoms and photons across distributed networks.

Our efforts follow the pioneering work of V. Braginsky *et al.*²⁴ and later studies²⁵ by employing the whispering-gallery modes of fused silica microtoroidal resonators²⁶. As shown in Fig. 1, a silicon chip containing a collection of 35 microtoroidal resonators is located

inside a vacuum chamber at 10^{-9} Torr and is positioned to couple a particular resonator to a tapered fibre. The toroids have major diameter $D \approx 44 \mu\text{m}$ and minor diameter $d \approx 6 \mu\text{m}$ (ref. 26). By judicious choice of the point of contact between the surface of the resonator and the tapered fibre, we attain critical coupling, in which the forward propagating power P_F in the fibre drops to near zero for the probe frequency ω_P equal to the cavity resonance frequency ω_C (ref. 23). Measurements of the cavity transmission in the absence of atoms are presented in Fig. 2. Note that the forward flux P_F and the associated transmission spectrum T_F are analogous to the reflected flux and reflection spectrum from a Fabry–Perot cavity²³. By varying the temperature of the silicon chip, the detuning $\Delta_{AC} \equiv \omega_C - \omega_A$ between ω_C and the atomic resonance at ω_A ($6S_{1/2}$; $F = 4 \rightarrow 6P_{3/2}$; $F' = 5'$ transition in caesium) can be controlled with an uncertainty of ± 2 MHz (Supplementary Information).

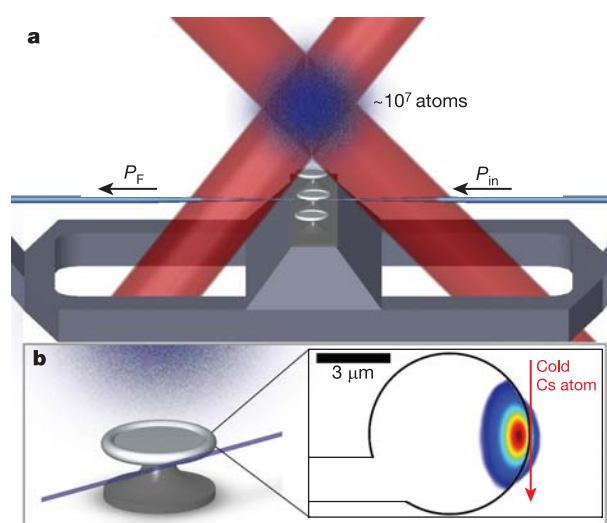


Figure 1 | Simple diagram of the experiment. **a**, A cloud of cold caesium atoms and the associated trapping lasers above an array of microtoroidal resonators. Light from the probe beam P_{in} is coupled into a resonator by way of the fibre taper, with the forward propagating output P_F coupled back into the taper from the resonator. **b**, Illustration of a SiO_2 microtoroidal resonator, fibre taper, and atom cloud above. The calculated field distribution for the lowest-order resonator mode is shown by the colour contour plot on the right. Cold caesium atoms fall through the external evanescent field of this mode and are thereby strongly coupled to the resonator's field.

¹Norman Bridge Laboratory of Physics 12-33, California Institute of Technology, Pasadena, California 91125, USA. ²T. J. Watson Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA. [†]Present addresses: Department of Applied Physics, The University of Tokyo, Tokyo 113-8656, Japan (T.A.); Physics Department, University of Otago, Dunedin 9016, New Zealand (W.P.B.); Department of Physics, University of Auckland, Auckland 1142, New Zealand (A.S.P.); Max Planck Institute of Quantum Optics, Garching 85748, Germany (T.J.K.).

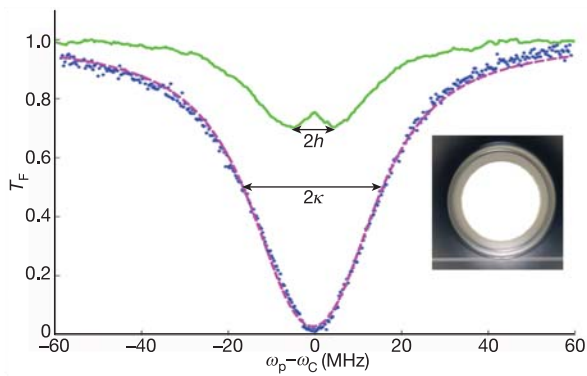


Figure 2 | Cavity transmission function $T_F = P_F/P_{in}$ as a function of probe frequency ω_p . The lower trace is taken for critical coupling, and the upper trace for conditions of under-coupling²³. From fits to such traces for critical coupling (red dashed curve), we find that $(\kappa, h)/2\pi = (17.9 \pm 2.8, 4.9 \pm 1.3)$ MHz, with κ, h being the overall cavity field decay rate and the scattering-induced coupling between the two counter-propagating modes of the microtoroid, respectively (see Supplementary Information for more details). Inset: photograph of a microtoroid and coupling fibre.

Cold atoms are delivered to the vicinity of the toroidal resonator from a small cloud of caesium atoms cooled to $T \approx 10 \mu\text{K}$ and located 10 mm above the silicon chip. Every 5 s, the cloud is dropped, resulting in about 2×10^6 atoms in a 3-mm ball at the height of the chip, with then a few dozen atoms passing through the external evanescent field of the toroidal resonator. By way of two single-photon detectors (D_{F1} , D_{F2}) (Supplementary Information), we continuously monitor the forward propagating signal P_F from a frequency-stabilized probe beam P_{in} coupled to the toroidal resonator. The interaction of an individual atom with the evanescent field destroys the condition of critical coupling, leading to an increase in

P_F . The measurement cycle then repeats itself for 2.5 s for a reference measurement, this time with no atomic cloud formed above the microtoroid.

Figure 3 displays typical records $C(t)$ for the number of single-photon detection events within time bins of $\delta t = 2 \mu\text{s}$ as functions of time t for the forward signal $P_F(t)$. Measurements are displayed with (Fig. 3a) and without (Fig. 3b) atoms for the case of equal probe and cavity frequencies, $\omega_p = \omega_c$, for $\Delta_{AC} \approx 0$, and with mean intracavity photon number $n_0 \approx 0.3$ for the forward circulating mode a of the toroidal resonator (Supplementary Information). The traces in both Fig. 3a and Fig. 3b exhibit background levels that result from the non-zero ratio $P_F/P_{in} \approx 0.005$ at critical coupling in the absence of atoms. However, Fig. 3a clearly shows sharp peaks of duration $\Delta t \approx 2 \mu\text{s}$ for the forward-propagating light $P_F(t)$, with an individual peak shown more clearly in the inset. Each event arises from the transit of a single atom through the resonant mode of the microtoroid, with about 30 events per cycle observed. Figure 3c examines the temporal profile of transit events in more detail by way of the cross-correlation $I(\tau)$ of photoelectric counts $C_1(t_1)$ and $C_2(t_1 + \tau)$ from the detectors D_{F1} and D_{F2} for P_F (Supplementary Information). This result agrees reasonably well with the theoretical prediction for atom transits through the calculated field distribution shown in Fig. 1b.

By applying a threshold requiring $C(t) \geq 6$ counts for $C(t)$ as in Fig. 3a, b, we find the average time dependence $\bar{C}_{\geq 6}(t)$ over about 100 measurement cycles. Figure 3d displays the results both with and without atoms, with the average counts $\sum_6(t)$ derived from $\bar{C}_{\geq 6}(t)$ by summing over successive time bins $\delta t = 2 \mu\text{s}$ for 1-ms intervals. The peak in transit events is consistent with the expected distribution of arrival times for atoms dropped from our atom cloud. By contrast, negligible excess events (that is, $C(t) \geq 6$) are recorded for the cases without atoms.

Focusing attention to the central region indicated by the dashed lines in Fig. 3d, we examine in Fig. 3e the probability $P(C)$ of recording C counts within $\delta t = 2 \mu\text{s}$. Evidently, when the atom

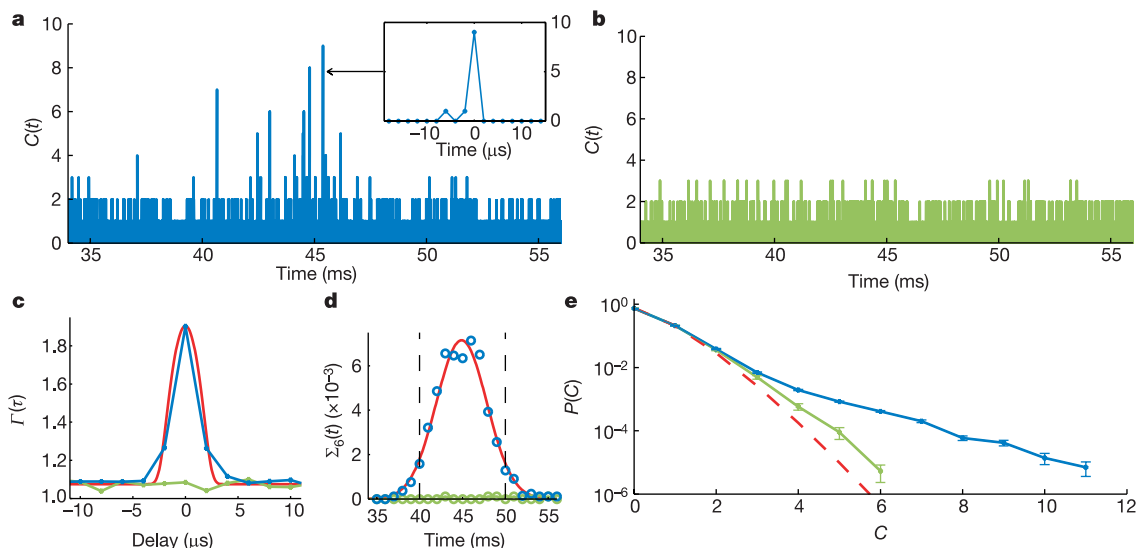


Figure 3 | Measurements of the forward signal P_F in the presence of falling atoms (blue) and without atoms (green). **a, b**, Single-photon counting events $C(t)$ as a function of time t after the release of the cold atom cloud at $t = 0$, with **(a)** and without **(b)** atoms dropped. $C(t)$ gives the total number of counts recorded for time bins of $\delta t = 2 \mu\text{s}$ duration. The inset in **a** shows the time profile for a single-atom transit. **c**, Normalized cross-correlation $I(\tau)$ of the forward signal counts from two detectors (D_{F1} , D_{F2}), showing the time profile associated with atom transit events. The smooth (red) curve is the theoretically predicted average cross-correlation for a transit event with one atom, taking into account the drop height of 10 mm and the spatial shape of the mode, as depicted in Fig. 1b. **d**, Counts $\sum_6(t)$ obtained from

$\bar{C}_{\geq 6}(t)$ by summing over 1-ms intervals, compared with a gaussian distribution that fits the rate of atom transits assuming a 3-mm (full-width at half-maximum) cloud of cold atoms dropped from 10 mm above the microtoroid. **e**, Probability $P(C)$ of detecting C counts within time bins of $\delta t = 2 \mu\text{s}$ for the central interval shown by the vertical dashed lines in **d**, compared with poissonian statistics (red) with the same mean number of counts (~ 0.25 per $2 \mu\text{s}$). The excess probability above the poissonian level in the no atoms case is predominantly due to instability in the cavity temperature, which results in small fluctuations in the forward flux. Error bars show ± 1 s.d.

cloud is present, there is a statistically significant increase (of at least 15σ) in the number of events with $C \geq 4$. These are precisely the events illustrated by the inset in Fig. 3a and the cross-correlation in Fig. 3c, and are associated with single-atom transits near the surface of the toroidal resonator. By varying the value of n_0 we have confirmed that the large transit events evident in Fig. 3 are markedly decreased for $n_0 \geq 1$ photons, which indicates the saturation of the atom–cavity system.

A quantitative description of our observations in Fig. 3 of individual atom transits requires the development of a new theoretical model in cavity QED. In the Supplementary Information we present such a model and show that the underlying description of the interaction of an atom with the fields of the toroidal resonator is in terms of normal modes (A,B) (Supplementary Fig. 1), which have mode functions $\psi_{A,B}(\rho, x, z)$ that are standing waves ($\cos kx$, $\sin kx$) around the circumference x of the toroid, with ρ the radial distance from the surface and z the vertical coordinate. Mode functions $\psi_{A,B}(\rho, x, z)$ have a calculated peak coherent coupling $g_0/2\pi$ of 70 MHz for the lowest-order modes of our resonator (such as that illustrated in Fig. 1b). The normal modes A,B result from the coupling of two oppositely directed travelling waves by scattering at rate h , with the resulting mode splitting seen in Fig. 2. Note that the presence of two normal modes leads to a $\sqrt{2}$ increase in the coupling constant in our case in contrast with that predicted by the Jaynes–Cummings model for an atom interacting with a single travelling-wave mode (see Supplementary Information for further details).

Guided by this theory, we have performed a series of measurements similar to those presented in Fig. 3 to determine the coherent coupling rate g_0 for interactions of single atoms with our toroidal resonator, but now with various values of the atom–cavity detuning Δ_{AC} , keeping the probe resonant with the cavity: $\omega_C \approx \omega_p = \omega_A + \Delta_{AC}$. The

qualitative idea is that large single-atom transit events will occur only over a range of detunings Δ_{AC} determined by g_0 . Specifically, the decrease in the forward transmission T_F induced by atom transits as a function of Δ_{AC} is described by a lorentzian with width β set by g_0 (Supplementary Information). In our case, $g_0 = g_0(\rho, x, z) \approx g_0(\rho, x, Vt)$, where V is the velocity of the dropped atoms in the z direction. Thus, a numerical integration was performed over ρ , x and t to derive the theoretical expectation for $T_F(\Delta_{AC})$, presented in Fig. 4a for three values of g_0^m , where g_0^m is the maximal coupling that an atom can experience in its interaction with the cavity modes. Indeed, we see that the width β grows monotonically with g_0^m . However, the average value of T_F is not readily measured in our current experiment, in which we expect many short individual transits, some of which are too weak to be distinguished from the background noise (see Fig. 3e). A parameter that describes our actual experimental measurements more closely is the probability of obtaining a transit that results in transmission above a certain threshold. The two measures are closely related, such that this probability decreases with detuning Δ_{AC} in the same fashion as T_F .

Figure 4b–d presents the results of our measurements for the average number of transit events per atom drop, $N_{\text{drop}}^{\text{av}}(C \geq C_0)$, which have photoelectric counts greater than or equal to a threshold value C_0 for a set of seven detunings Δ_{AC} . In accord with the expectation set by Fig. 4a, there is a decrease in the occurrence of large transit events for increasing Δ_{AC} in correspondence to the decrease in the effective atom–cavity coupling coefficient for large atom–cavity detunings. The full curves shown in Fig. 4b–d are the results of theoretical calculation for these measurements, with the relevant cavity parameters (κ , h) determined from fits as in Fig. 2.

We first ask whether the data might be explained by an effective value g_0^e for the coherent coupling of atom and cavity field, without taking into account the fact that individual atoms transit at radial distances ρ that vary from atom to atom. Figure 4b examines this possibility for various values of g_0^e , assuming a coupling coefficient $g_0^e \psi_{A,B}(x) = g_0^e [\cos kx, \sin kx]$, averaged along one period in x (as in Fig. 4a). Apparently, an effective value $g_0^e/2\pi = 40$ MHz provides reasonable correspondence between theory and experiment for large events $C \geq 6$.

We adapt our theory to the actual situation of atoms arriving randomly at radial and circumferential coordinates by introducing a mesh grid over (ρ, x) , and then computing the cavity transmission function $T_F(t)$ from $\psi_{A,B}(\rho, x, z(t))$ for atomic trajectories $z(t)$ over this grid. We account for the time resolution $\delta t = 2 \mu\text{s}$ of our data acquisition by a suitable average of $T_F(t)$ over such time bins (as was also true in Fig. 4b). The results from these calculations are shown in Fig. 4c, d as the set of full curves for three values of coherent coupling g_0 for the cavity mode functions $\psi_{A,B}(\rho, x, z)$, where in Fig. 4b–d the theory is scaled to match the measured $N_{\text{drop}}^{\text{av}}(C \geq C_0)$ at $\Delta_{AC} = 0$. From such comparisons we determine a maximal accessible $g_0^m/2\pi$ of 50 ± 12 MHz. This conclusion is insensitive to the choice of cut-off C_0 over the range $4 \leq C_0 \leq 9$ for which we have significant transit events. Strong coupling with $g_0^m > (\kappa, \gamma)$ is thereby achieved, where $(\kappa, \gamma)/2\pi = (17.9 \pm 2.8, 2.6)$ MHz are the dissipative rates for the cavity field and the atom.

According to our calculations, $g_0^m/2\pi = 50$ MHz corresponds to the coupling rate expected at a distance of roughly 45 nm from the surface of the microtoroid. We estimate that due to the attractive van der Waals forces²⁷, atoms which enter the evanescent field with a distance $\rho \leq 45$ nm from the microtoroid are expected to strike its surface in less than $1 \mu\text{s}$, thus preventing such atoms from generating appreciable transit events in the transmission function T_F .

Thus, we report strong coupling for single atoms interacting with an optical resonator other than a conventional Fabry–Perot cavity. The monolithic microtoroidal resonators²² employed here have the capability of input–output coupling with small parasitic losses, with a demonstrated ideality of more than 99.97%²³. Moreover, quality factors $Q = 4 \times 10^8$ have been realized at $\lambda = 1,550$ nm (ref. 28) and

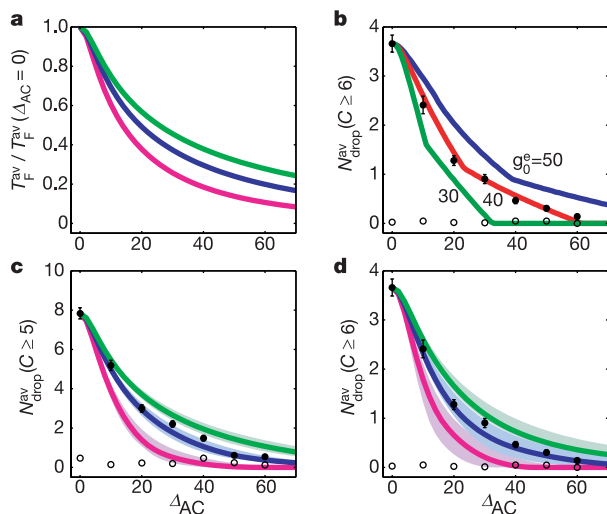


Figure 4 | Measurements of transit events as a function of the atom–cavity detuning Δ_{AC} . Events are shown in the presence of atoms (filled circles) and without atoms (empty circles), compared with the theoretical calculations (lines). **a**, Theoretical calculation for the average of the transmission $T_F(\omega_p = \omega_C)$ as a function of (Δ_{AC}, g_0) . Red, $g_0 = 35$; blue, $g_0 = 50$; green, $g_0 = 65$. **b–d**, Measurements of the average number of events per drop of the atom cloud $N_{\text{drop}}^{\text{av}}(C \geq C_0)$ plotted against the atom–cavity detuning Δ_{AC} , with $C_0 = 6$ (**b**, **d**) and $C_0 = 5$ (**c**). Error bars show ± 1 s.d. The data are taken for a cavity resonance equal to the probe frequency $\omega_C \approx \omega_p = \omega_A + \Delta_{AC}$. The full curves are theoretical results as discussed in the text. The widths of the curves are determined from the experimental uncertainties in (κ, h) . **b**, Theory for $N_{\text{drop}}^{\text{av}}(C \geq 6)$ without radial averaging to deduce an effective coupling $g_0^e/2\pi = 40$ MHz. Green, $g_0^e = 30$; red, $g_0^e = 40$; blue, $g_0^e = 50$. **c**, **d**, Theory for $N_{\text{drop}}^{\text{av}}(C \geq 5)$, $N_{\text{drop}}^{\text{av}}(C \geq 6)$, respectively, with radial and azimuthal averaging leading to $g_0^m/2\pi = 50$ MHz. Red, $g_0^m = 35$; blue, $g_0^m = 50$; green, $g_0^m = 65$.

$Q \approx 10^8$ at $\lambda = 850$ nm (ref. 26), with good prospects for improvements to $Q \approx 10^{10}$ (ref. 29). For these parameters, the efficiency for coupling single photons into and out of the resonator could approach $\varepsilon \sim 0.99$ – 0.999 (ref. 23), while still remaining firmly in the regime of strong coupling²⁶. Such high efficiency is critical for the realization of scalable quantum networks^{18,19} and photonic quantum computation²⁰. Indeed, of the diverse possibilities for the distribution and processing of quantum information with optical cavities^{5,7,8}, the system of single atoms coupled to microtoroidal resonators arguably provides one of the most promising avenues. Beyond efficient input–output coupling²³, strong coupling to a material system with long-lived internal states has now been showed, although here in a primitive, proof-of-principle setting. An outstanding technical challenge is to trap single atoms near the surface of the microtoroid, with one possibility having been investigated in ref. 30.

Received 2 June; accepted 1 August 2006.

1. Miller, R. *et al.* Trapped atoms in cavity QED: coupling quantized light and matter. *J. Phys. B At. Mol. Opt. Phys.* **38**, S551–S565 (2005).
2. Berman, P. (ed.) *Cavity Quantum Electrodynamics* (Academic, San Diego, 1994).
3. Walther, H. Quantum optics of single atoms. *Fortschr. Phys.* **52**, 1154–1164 (2004).
4. Raimond, J. M. *et al.* Probing a quantum field in a photon box. *J. Phys. B At. Mol. Opt. Phys.* **38**, S535–S550 (2005).
5. Vahala, K. J. Optical microcavities. *Nature* **424**, 839–846 (2003).
6. Nussmann, S. *et al.* Vacuum-stimulated cooling of single atoms in three dimensions. *Nature Phys.* **1**, 122–125 (2005).
7. Khitrova, G., Gibbs, H. M., Kira, M., Koch, S. W. & Scherer, A. Vacuum Rabi splitting in semiconductors. *Nature Phys.* **2**, 81–90 (2006).
8. Badolato, A. *et al.* Deterministic coupling of single quantum dots to single nanocavity nodes. *Science* **308**, 1158–1161 (2005).
9. Wallraff, A. *et al.* Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431**, 162–167 (2004).
10. Chiorescu, I. *et al.* Coherent dynamics of a flux qubit coupled to a harmonic oscillator. *Nature* **431**, 159–162 (2004).
11. McKeever, J., Boca, A., Boozer, A. D., Buck, J. R. & Kimble, H. J. Experimental realization of a one-atom laser in the regime of strong coupling. *Nature* **425**, 268–271 (2003).
12. McKeever, J. *et al.* Deterministic generation of single photons from one atom trapped in a cavity. *Science* **303**, 1992–1994 (2004).
13. Keller, M., Lange, B., Hayasaka, K., Lange, W. & Walther, H. Continuous generation of single photons with controlled waveform in an ion-trap cavity system. *Nature* **431**, 1075–1078 (2004).
14. Birnbaum, K. M. *et al.* Photon blockade in an optical cavity with one trapped atom. *Nature* **436**, 87–90 (2005).
15. Legero, T., Wilk, T., Hennrich, M., Rempe, G. & Kuhn, A. Quantum beat of two single photons. *Phys. Rev. Lett.* **93**, 070503 (2004).
16. Rempe, G., Thompson, R. J., Kimble, H. J. & Lalezari, R. Measurement of ultralow losses in an optical interferometer. *Opt. Lett.* **17**, 363–365 (1992).
17. Hood, C. J., Ye, J. & Kimble, H. J. Characterization of high-finesse mirrors: loss, phase shifts, and mode structure in an optical cavity. *Phys. Rev. A* **64**, 033804 (2001).
18. Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).
19. Briegel, H.-J., van Enk, S. J., Cirac, J. I. & Zoller, P. in *The Physics of Quantum Information* (eds Bouwmeester, D., Ekert, A. & Zeilinger, A.) 192–197 (Springer, Berlin, 2000).
20. Duan, L.-M. & Kimble, H. J. Scalable photonic quantum computation through cavity-assisted interactions. *Phys. Rev. Lett.* **92**, 127902 (2004).
21. Treutlein, P. *et al.* Quantum information processing in optical lattices and magnetic microtraps. Preprint at (<http://arxiv.org/abs/quant-ph/0605163>) (2006).
22. Armani, D. K., Kippenberg, T. J., Spillane, S. M. & Vahala, K. J. Ultra-high-Q toroid microcavity on a chip. *Nature* **421**, 925–928 (2003).
23. Spillane, S. M., Kippenberg, T. J., Painter, O. J. & Vahala, K. J. Ideality in a fiber-taper-coupled microresonator system for application to cavity quantum electrodynamics. *Phys. Rev. Lett.* **91**, 043902 (2003).
24. Braginsky, V. B., Gorodetsky, M. L. & Ilchenko, V. S. Quality-factor and nonlinear properties of optical whispering-gallery modes. *Phys. Lett. A* **137**, 393–397 (1989).
25. Vernooy, D. W., Furusawa, A., Georgiades, N. Ph., Ilchenko, V. S. & Kimble, H. J. Cavity QED with high-Q whispering gallery modes. *Phys. Rev. A* **57**, R2293–R2296 (1998).
26. Spillane, S. M. *et al.* Ultrahigh-Q toroidal microresonators for cavity quantum electrodynamics. *Phys. Rev. A* **71**, 013817 (2005).
27. Courtois, J.-Y., Courty, J.-M. & Mertz, J. C. Internal dynamics of multilevel atoms near a vacuum-dielectric interface. *Phys. Rev. A* **53**, 1862–1878 (1996).
28. Kippenberg, T. J., Spillane, S. M. & Vahala, K. J. Demonstration of ultra-high-Q small mode volume toroid microcavities on a chip. *Appl. Phys. Lett.* **85**, 6113–6115 (2004).
29. Vernooy, D. W., Ilchenko, V. S., Mabuchi, H., Streed, E. W. & Kimble, H. J. High-Q measurements of fused-silica microspheres in the near infrared. *Opt. Lett.* **23**, 247–249 (1998).
30. Vernooy, D. W. & Kimble, H. J. Quantum structure and dynamics for atom galleries. *Phys. Rev. A* **55**, 1239–1261 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Eichenfield, K. W. Goh and S. M. Spillane for their contributions to the early stages of this experiment, and T. Carmon, A. Gross and S. Walavalkar for their contributions to the current realization. The work of H.J.K. is supported by the National Science Foundation, the Disruptive Technology Office of the Department of National Intelligence, and Caltech. The work of K.J.V. is supported by DARPA, the Caltech Lee Center and the National Science Foundation. B.D., W.P.B. and T.J.K. acknowledge support as Fellows of the Center for the Physics of Information at Caltech. A.S.P. acknowledges support from the Marsden Fund of the Royal Society of New Zealand. E.W. acknowledges support as a Ford Predoctoral Fellow from the US National Academies.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.J.K. (hjkimble@caltech.edu).

The order of the quantum chromodynamics transition predicted by the standard model of particle physics

Y. Aoki¹, G. Endrődi², Z. Fodor^{1,2}, S. D. Katz^{1,2} & K. K. Szabó¹

Quantum chromodynamics (QCD) is the theory of the strong interaction, explaining (for example) the binding of three almost massless quarks into a much heavier proton or neutron—and thus most of the mass of the visible Universe. The standard model of particle physics predicts a QCD-related transition that is relevant for the evolution of the early Universe. At low temperatures, the dominant degrees of freedom are colourless bound states of hadrons (such as protons and pions). However, QCD is asymptotically free, meaning that at high energies or temperatures the interaction gets weaker and weaker^{1,2}, causing hadrons to break up. This behaviour underlies the predicted cosmological transition between the low-temperature hadronic phase and a high-temperature quark–gluon plasma phase (for simplicity, we use the word ‘phase’ to characterize regions with different dominant degrees of freedom). Despite enormous theoretical effort, the nature of this finite-temperature QCD transition (that is, first-order, second-order or analytic crossover) remains ambiguous. Here we determine the nature of the QCD transition using computationally demanding lattice calculations for physical quark masses. Susceptibilities are extrapolated to vanishing lattice spacing for three physical volumes, the smallest and largest of which differ by a factor of five. This ensures that a true transition should result in a dramatic increase of the susceptibilities. No such behaviour is observed: our finite-size scaling analysis shows that the finite-temperature QCD transition in the hot early Universe was not a real phase transition, but an analytic crossover (involving a rapid change, as opposed to a jump, as the temperature varied). As such, it will be difficult to find experimental evidence of this transition from astronomical observations.

During the evolution of the Universe there were particle-physics-related transitions. Although there are strong indications of an inflationary period, we know little about how it affected possible transitions of our known physical model. To understand the consequences, we need a clear picture about these cosmologically relevant transitions. The Standard Model (SM) of particle physics predicts two such transitions.

One of the SM-based transitions occurs at temperatures (T) of a few hundred GeV. This transition is responsible for the spontaneous breaking of the electroweak symmetry, which gives the masses of the elementary particles. This transition is also related to the electroweak baryon-number violating processes, which had a major influence on the observed baryon-asymmetry of the Universe. Lattice results have shown that the electroweak transition in the SM is an analytic crossover^{3–6}.

The second transition occurs at $T \approx 200$ MeV. It is related to the spontaneous breaking of the chiral symmetry of QCD. The nature of the QCD transition affects our understanding of the Universe’s

evolution (see ref. 7 for example). In a strong first-order phase transition the quark–gluon plasma supercools before bubbles of hadron gas are formed. These bubbles grow, collide and merge, during which gravitational waves could be produced⁸. Baryon-enriched nuggets could remain between the bubbles, contributing to dark matter. The hadronic phase is the initial condition for nucleosynthesis, so inhomogeneities in this phase could have a strong effect on nucleosynthesis⁹. As the first-order phase transition weakens, these effects become less pronounced. Our calculations provide strong evidence that the QCD transition is a crossover, and thus the above scenarios—and many others—are ruled out.

We emphasize that extensive experimental work is currently being done with heavy ion collisions to study the QCD transition (most recently at the Relativistic Heavy Ion Collider, RHIC). Both for the cosmological transition and for RHIC, the net baryon densities are quite small, and so the baryonic chemical potentials (μ) are much less than the typical hadron masses (~ 45 MeV at RHIC and negligible in the early Universe). A calculation at $\mu = 0$ is directly applicable for the cosmological transition and most probably also determines the nature of the transition at RHIC. Thus we carry out our analysis at $\mu = 0$.

QCD is a generalized version of quantum electrodynamics (QED). The euclidean lagrangian with gauge coupling g and with a quark mass of m can be written as $\mathcal{L} = -1/(2g^2)\text{Tr}F_{\mu\nu}F_{\mu\nu} + \bar{\psi}\gamma_\mu(\partial_\mu + A_\mu + m)\psi$, where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$. In electrodynamics the gauge

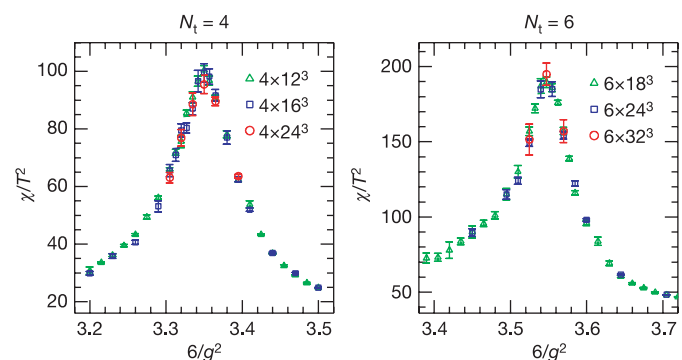


Figure 1 | Susceptibilities for the light quarks for $N_t = 4$ and for $N_t = 6$ as a function of $6/g^2$, where g is the gauge coupling. (T grows with $6/g^2$.) The largest volume is eight times bigger than the smallest one, so a first-order phase transition would predict a susceptibility peak that is eight times higher (for a second-order phase transition the increase would be somewhat less, but still dramatic). Instead of such a significant change we do not observe any volume dependence. Error bars are s.e.m.

¹Department of Physics, University of Wuppertal, D-42097 Wuppertal, Germany. ²Institute for Theoretical Physics, Eötvös University, H-1117 Budapest, Hungary.

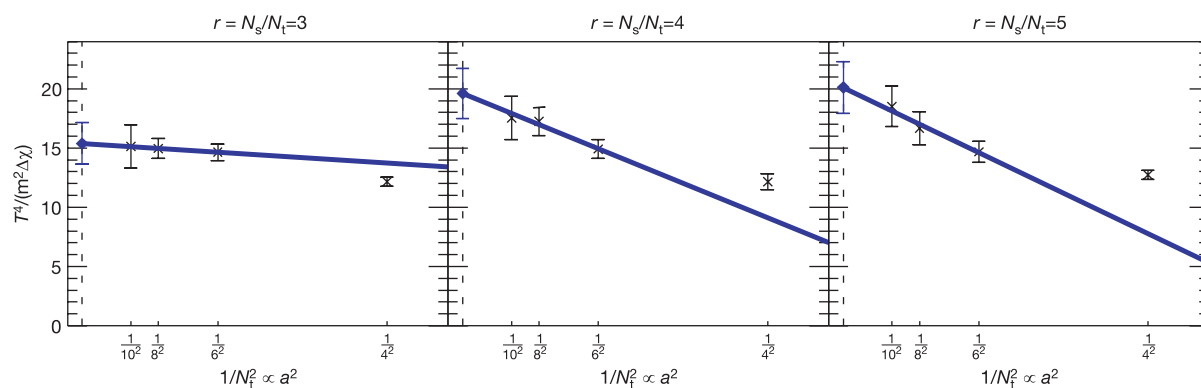


Figure 2 | Normalized susceptibilities $T^4/(m^2\Delta\chi)$ for the light quarks for given aspect ratios (r) as functions of the lattice spacing. Continuum extrapolations are carried out for all three physical volumes and

the results are given by the leftmost blue diamonds. Error bars are s.e.m. with systematic estimates.

field A_μ is a simple real field, whereas in QCD it is a 3×3 matrix. Consequently the commutator in $F_{\mu\nu}$ vanishes for QED, but does not vanish in QCD. The ψ fields also have an additional 'colour' index in QCD, which runs from 1 to 3. Different types of quarks are represented by fermionic fields with different masses. The action S is defined as the four-volume integral of $\mathcal{L}$. The basic quantity we determine is the partition function Z , which is the sum of the Boltzmann factors $\exp(-S)$ for all field configurations. Partial derivatives of Z with respect to the masses give rise to the order parameters we studied here.

There are some QCD results and model calculations to determine the order of the transition at $\mu = 0$ and $\mu \neq 0$ for different fermionic contents (compare refs 10–19). Unfortunately, none of these approaches can give an unambiguous answer for the order of the transition for physical values of the quark masses. The only known systematic technique which could give a final answer is lattice QCD.

Lattice QCD discretizes the above lagrangian on a four-dimensional lattice and extrapolates the results to vanishing lattice spacing ($a \rightarrow 0$). A convenient way to carry out this discretization is to put the fermionic variables on the sites of the lattice, whereas the gauge fields are treated as 3×3 matrices connecting these sites. In this sense, lattice QCD is a classical four-dimensional statistical physics system. One important difference (compared to three-dimensional systems) is that the temperature (T) is determined by the additional, euclidean time extension (N_t): $T = 1/(N_t a)$. Keeping the temperature fixed (such as at the transition point) one can reduce a and approach the continuum limit by increasing N_t (see Methods).

There are several lattice results for the order of the QCD transition (for the two most popular lattice fermion formulations see refs 20 and 21), although they have unknown systematics. We emphasize that from the lattice point of view two 'ingredients' are necessary to eliminate these systematic uncertainties.

The first ingredient is to use physical values for the quark masses. Owing to the computational costs this is a great challenge in lattice QCD. Previous analyses used computationally less-demanding non-physically large quark masses. However, these choices have limited relevance. The order of the transition depends on the quark mass. For example, in three-flavour QCD for vanishing quark masses the transition is of first order. For intermediate masses it is probably a crossover. For infinitely heavy quark masses the transition is again first order. For questions concerning the restoration of chiral symmetry (such as the order of the transition), a controlled extrapolation from larger quark masses (such as chiral perturbation theory) is unavailable, and so the physical quark masses should be used directly.

The second ingredient is to remove the uncertainty associated with the lattice discretization. Discretization errors disappear in the continuum limit; however, they strongly influence the results at non-vanishing lattice spacing. In three-flavour unimproved staggered

QCD, using a lattice spacing of about 0.28 fm, the first-order and the crossover regions are separated by a pseudoscalar mass of $m_{\pi,c} \approx 300$ MeV. Studying the same three-flavour theory with the same lattice spacing, but with an improved p4 action (which has different discretization errors) we obtain $m_{\pi,c} \approx 70$ MeV. In the first approximation, a pseudoscalar mass of 140 MeV (which corresponds to the numerical value of the physical pion mass) would be in the first-order transition region, whereas using the second approximation, it would be in the crossover region. The different discretization uncertainties are solely responsible for these qualitatively different results²². Therefore, the proper approach is to use physical quark masses, and to extrapolate to vanishing lattice spacings. Our work eliminates both of the above uncertainties.

Our goal is to identify the nature of the transition for physical quark masses as we approach the continuum limit. We will study the finite size scaling of the lattice chiral susceptibilities $\chi(N_s, N_t) = \partial^2/(\partial m_{ud}^2)(T/V) \log Z$, where m_{ud} is the mass of the light u, d quarks and N_s is the spatial extension. This susceptibility shows a pronounced peak around the transition temperature (T_c). For a real phase transition the height of the susceptibility peak increases and the width of the peak decreases when we increase the volume. For a first-order phase transition the finite size scaling is determined by the geometric dimension, the height is proportional

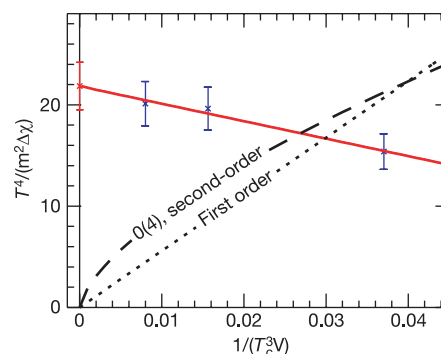


Figure 3 | Continuum extrapolated susceptibilities $T^4/(m^2\Delta\chi)$ as a function of $1/(T_c^3 V)$. For true phase transitions the infinite volume extrapolation should be consistent with zero, whereas for an analytic crossover the infinite volume extrapolation gives a non-vanishing value. The continuum-extrapolated susceptibilities show no phase-transition-like volume dependence, though the volume changes by a factor of five. The $V \rightarrow \infty$ extrapolated value is 22(2), which is 11σ away from zero. For illustration, we fit the expected asymptotic behaviour for first-order and O(4) (second order) phase transitions shown by dotted and dashed lines, which results in chance probabilities of 10^{-19} and 7×10^{-13} , respectively. Error bars are s.e.m. with systematic estimates.

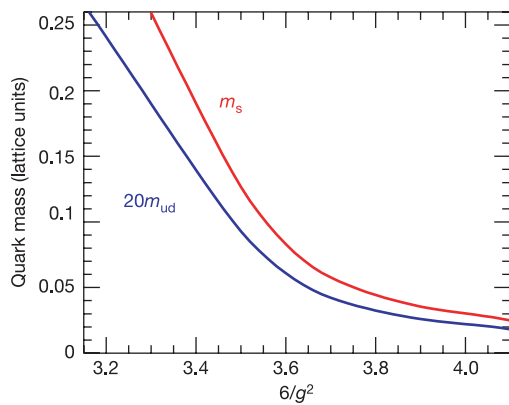


Figure 4 | The line of constant physics. We show our choice for m_s (strange quark mass) and $20m_{ud}$ (u, d quark masses) in lattice units as functions of $6/g^2$.

to V , and the width is proportional to $1/V$. For a second-order transition the singular behaviour is given by some power of V , defined by the critical exponents. The picture would be completely different for an analytic crossover. There would be no singular behaviour and the susceptibility peak would not get sharper when we increase the volume; instead, its height and width will be V independent for large volumes.

Figure 1 shows the susceptibilities for the light quarks for $N_t = 4$ and 6, for which we used aspect ratios $r = N_s/N_t$ ranging from 3 to 6 and 3 to 5, respectively. A clear signal for an analytic crossover for both lattice spacings can be seen. However, these curves do not say much about the continuum behaviour of the theory. In principle, a phenomenon as unfortunate as that in the three-flavour theory could occur²², in which the reduction of the discretization effects changed the nature of the transition for a pseudoscalar mass of ~ 140 MeV.

Because we are interested in genuine temperature effects we subtract the $T = 0$ susceptibility and study only the difference between $T \neq 0$ and $T = 0$ at different lattice spacings. To do it properly, when we approach the continuum limit the renormalization of χ has to be performed. This leads to $m^2\Delta\chi$, which we study (see Methods).

To give a continuum result for the order of the transition we carry out a finite size scaling analysis of the dimensionless quantity $T^4/(m^2\Delta\chi)$ directly in the continuum limit. For this study we need the height of the susceptibility peaks in the continuum limit for fixed physical volumes. The continuum extrapolations are done using four different lattice spacings ($N_t = 4, 6, 8$ and 10). The volumes at different lattice spacings are fixed in units of T_c , and thus $VT_c^3 = 3^3, 4^3$ and 5^3 were chosen. (In three cases the computer architecture did not allow us to take the above ratios directly. In these cases, we used the next possible volume and interpolated or extrapolated. The height of the peak depends weakly on the volume, so these procedures were always safe.) Altogether, we used twelve different lattice volumes ranging from 4×12^3 to 10×48^3 at $T > 0$. For the $T = 0$ runs lattice volumes from 24×12^3 up to 56×28^3 were used. The number of trajectories were between 1,500 and 8,000 for $T > 0$ and between 1,500 and 3,000 for $T = 0$, respectively. Figure 2 shows the continuum extrapolation for the three different physical volumes. The $N_t = 4$ results are slightly off but the $N_t = 6, 8$ and 10 results show a good $a^2 \propto 1/N_t^2$ scaling.

Having obtained the continuum values for $T^4/(m^2\Delta\chi)$ at fixed physical volumes, we study the finite size scaling of the results. Figure 3 shows our final results. The volume dependence strongly suggests that there is no true phase transition but only an analytic crossover in QCD.

METHODS

An introduction to lattice QCD at $T = 0$ and $T \neq 0$ can be found in refs 23 and 24, for example. The detailed form of our Symanzik improved gauge and stout-link improved staggered fermionic action can be found in ref. 25. (Staggered QCD with $n_f = 4$ flavours uses the fourth-root description for the fermionic determinant, the relevance of which was recently intensively discussed, see ref. 26 and references therein.) In this work we also used a stout-smearing level of 2. Note that stout-link improvement makes the staggered fermion taste symmetry violation small even at moderate lattice spacings (for an illustration, see Fig. 1 of ref. 25). Taste symmetry violation errors—as with other discretization errors of the staggered formalism—scale as a^2 and disappear only after extrapolating to the continuum limit. We carried out this extrapolation and explicitly checked that taste symmetry violations for our smallest lattice spacings are already within this a^2 scaling regime, and so we conclude that the extrapolation is reliable.

In previous staggered analyses the gauge configurations were produced by the R -algorithm²⁷ at a given step size. The step size is an intrinsic parameter of the algorithm, which has to be extrapolated to zero. Instead of using the approximate R -algorithm, we used²⁵ the exact rational hybrid Monte-Carlo algorithm²⁸ in large-scale simulations. This technique, which we apply also in this paper, expresses the fractional powers of the Dirac operator by rational functions.

In our simulations we approach the continuum limit along the line of constant physics (LCP). The LCP is defined by relationships between the bare lattice parameters (gauge coupling g and lattice bare quark masses for light quarks m_{ud} and strange quark m_s). These relationships ensure that the physics (such as ratios of physical observables) remain constant while changing any of the parameters. We note that the LCP is unambiguous (independent of the physical quantities, which are used to define the above relationships) only in the vicinity of the continuum limit. The present work uses the LCP defined by fixing two ratios to their physical values: $m_K/m_\pi = 3.689$ and $f_K/m_\pi = 1.185$ (where m_K is the mass of the kaon, f_K is its decay constant and m_π is the mass of the pion). Figure 4 shows the LCP used in this work.

Using the above action, simulation algorithm and parameter set based on our LCP, we performed simulations at $T \neq 0$ and $T = 0$. At $T \neq 0$ we always used the physical quark masses at four different temporal extensions: $N_t = 4, 6, 8$ and 10 lattices. The aspect ratios (N_s/N_t) were taken to be 3, 4 and 5 (for the $N_t = 4$ case also 6), which results in a factor-of-8 change in the volume (V). In the chirally broken phase (our zero-temperature simulations always belong to this class) chiral perturbation theory can be used to extrapolate to the physical values of the light quark masses in a controlled manner. Therefore we used four pion masses ($m_\pi \approx 235, 300, 355$ and 405 MeV) that are somewhat larger than the physical one. The spatial extension was chosen to ensure $N_s m_\pi \geq 4$. The computational requirement for the present work was 0.8 teraflopyears.

The renormalization of the chiral susceptibility can be done by taking the second derivative of the free energy density (f) with respect to the renormalized mass (m_r). We apply the usual definition: $f/T^4 = -N_t^4[\log Z(N_s, N_t)/(N_s N_t^3) - \log Z(N_{s0}, N_{t0})/(N_{s0} N_{t0}^3)]$. This quantity has a correct continuum limit. The subtraction term is obtained at $T = 0$, for which simulations are carried out on lattices with N_{s0}, N_{t0} spatial and temporal extensions (or otherwise at the same parameters of the action). The bare light quark mass is related to m_r by the mass renormalization constant: $m_r = Z_m m_{ud}$. Note that Z_m falls out of the combination $m_r^2 \partial^2 / \partial m_r^2 = m_{ud}^2 \partial^2 / \partial m_{ud}^2$. Thus, $m_{ud}^2[\chi(N_s, N_t) - \chi(N_{s0}, N_{t0})]$ also has a continuum limit (for its maximum values for different N_t , and in the continuum limit we use the shorthand notation $m^2\Delta\chi$). For the subtraction, $T = 0$ simulations are performed and chiral perturbation theory is applied.

Received 6 June; accepted 27 July 2006.

- Gross, D. J. & Wilczek, F. Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.* **30**, 1343–1346 (1973).
- Politzer, H. D. Reliable perturbative results for strong interactions? *Phys. Rev. Lett.* **30**, 1346–1349 (1973).
- Karsch, F., Neuhaus, T., Patkos, A. & Rank, J. Critical Higgs mass and temperature dependence of gauge boson masses in the SU(2) gauge-Higgs model. *Nucl. Phys. Proc. (Suppl.)* **53**, 623–625 (1997).
- Kajantie, K., Laine, M., Rummukainen, K. & Shaposhnikov, M. E. Is there a hot electroweak phase transition at $m_h > m_w$? *Phys. Rev. Lett.* **77**, 2887–2890 (1996).
- Csikor, F., Fodor, Z. & Heitger, J. Endpoint of the hot electroweak phase transition. *Phys. Rev. Lett.* **82**, 21–24 (1999).
- Gurtler, M., Ilgenfritz, E.-M. & Schiller, A. Where the electroweak phase transition ends. *Phys. Rev. D* **56**, 3888–3895 (1997).
- Schwarz, D. J. The first second of the universe. *Ann. Phys.* **12**, 220–270 (2003).
- Witten, E. Cosmic separation of phases. *Phys. Rev. D* **30**, 272–285 (1984).
- Applegate, J. H. & Hogan, C. J. Relics of cosmic quark condensation. *Phys. Rev. D* **31**, 3037–3045 (1985).

10. Pisarski, R. D. & Wilczek, F. Remarks on the chiral phase transition in chromodynamics. *Phys. Rev. D* **29**, 338–341 (1984).
11. Celik, T., Engels, J. & Satz, H. The order of the deconfinement transition in SU(3) Yang–Mills theory. *Phys. Lett. B* **125**, 411–414 (1983).
12. Kogut, J. B. *et al.* Deconfinement and chiral symmetry restoration at finite temperatures in SU(2) and SU(3) gauge theories. *Phys. Rev. Lett.* **50**, 393–396 (1983).
13. Gottlieb, S. A. *et al.* The deconfining phase transition and the continuum limit of lattice quantum chromodynamics. *Phys. Rev. Lett.* **55**, 1958–1961 (1985).
14. Brown, F. R., Christ, N. H., Deng, Y. F., Gao, M. S. & Woch, T. J. Nature of the deconfining phase transition in SU(3) lattice gauge theory. *Phys. Rev. Lett.* **61**, 2058–2061 (1988).
15. Fukugita, M., Okawa, M. & Ukawa, A. Order of the deconfining phase transition in SU(3) lattice gauge theory. *Phys. Rev. Lett.* **63**, 1768–1771 (1989).
16. Halasz, M. A., Jackson, A. D., Shrock, R. E., Stephanov, M. A. & Verbaarschot, J. J. M. On the phase diagram of QCD. *Phys. Rev. D* **58**, 096007 (1998).
17. Berges, J. & Rajagopal, K. Color superconductivity and chiral symmetry restoration at nonzero baryon density and temperature. *Nucl. Phys. B* **538**, 215–232 (1999).
18. Schaefer, B.-J. & Wambach, J. The phase diagram of the quark meson model. *Nucl. Phys. A* **757**, 479–492 (2005).
19. Herpay, T., Patkos, A., Szep, Zs. & Szeplalusy, P. Mapping the boundary of the first order finite temperature restoration of chiral symmetry in the m_π – m_K plane with a linear σ model. *Phys. Rev. D* **71**, 125017 (2005).
20. Brown, F. R. *et al.* On the existence of a phase transition for QCD with three light quarks. *Phys. Rev. Lett.* **65**, 2491–2494 (1990).
21. Khan, A. A. *et al.* Phase structure and critical temperature of two flavor QCD with renormalization group improved gauge action and clover improved Wilson quark action. *Phys. Rev. D* **63**, 034502 (2001).
22. Karsch, F. *et al.* Where is the chiral critical point in three-flavor QCD? *Nucl. Phys. Proc. Suppl.* **129**, 614–616 (2004).
23. Davies, C. T. H. Lattice QCD. In *Heavy Flavor Physics, Scottish Graduate Textbook Series* (eds Davies, C. T. H. & Playfer, S. M.) 7–23 (Institute of Physics, 2002).
24. Ukawa, A. Lectures on lattice qcd at finite temperature. In *Proceedings of the 1993 Uehling Summer School 'Phenomenology and Lattice QCD (Seattle, July 1993)'* (eds Kilcup, G. & Sharp, S.) (World Scientific, Singapore, 1995).
25. Aoki, Y., Fodor, Z., Katz, S. D. & Szabo, K. K. The equation of state in lattice QCD: with physical quark masses towards the continuum limit. *J. High-Energy Phys.* **01**:089 (2006).
26. Bernard, C., Golterman, M. & Shamir, Y. Observations on staggered fermions at non-zero lattice spacing. *Phys. Rev. D* **73**, 114511 (2006).
27. Gottlieb, S. A. *et al.* Hybrid molecular dynamics algorithms for the numerical simulation of quantum chromodynamics. *Phys. Rev. D* **35**, 2531–2542 (1987).
28. Clark, M. A. & Kennedy, A. D. The rhmc algorithm for two flavors of dynamical staggered fermions. *Nucl. Phys. Proc. Suppl.* **129**, 850–852 (2004).
29. Fodor, Z., Katz, S. D. & Papp, G. Better than \$1/mflops sustained: A scalable PC-based parallel computer for lattice qcd. *Comput. Phys. Commun.* **152**, 121–134 (2003).

Acknowledgements We thank F. Csikor, A. Dougall, K.-H. Kampert, M. Nagy, Z. Rácz and D. J. Schwarz for discussions. This research was partially supported by a DFG German Science grant, OTKA Hungarian Science grants and an EU research grant. The computations were carried out on PC clusters at the University of Budapest and Wuppertal with next-neighbour communication architecture²⁹ and on the BlueGene/L machine in Jülich. A modified version of the publicly available MILC code (<http://physics.indiana.edu/~sg/milc.html>) was used.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.D.K. (katz@bodri.elte.hu).

Origin of the dielectric dead layer in nanoscale capacitors

Massimiliano Stengel¹ & Nicola A. Spaldin¹

Capacitors are a mainstay of electronic integrated circuits and devices, where they perform essential functions such as storing electrical charge, and blocking direct current while allowing alternating currents to propagate. Because they are often the largest components in circuits, extensive efforts are directed at reducing their size through the use of high-permittivity insulators such as perovskite-structure SrTiO₃ (refs 1, 2), which should provide more capacitance per unit area of device. Unfortunately, most experiments on thin-film SrTiO₃ capacitors have yielded capacitance values that are orders of magnitude smaller than expected³. The microscopic origin of this reduced capacitance, which is often discussed in terms of a low-permittivity interfacial 'dead layer'⁴, is not well understood. Whether such a dead layer exists at all, and if so, whether it is an intrinsic property of an ideal metal–insulator interface or a result of processing issues such as defects and strains, are controversial questions. Here we present fully *ab initio* calculations of the dielectric properties of realistic SrRuO₃/SrTiO₃/SrRuO₃ nanocapacitors, and show that the observed dramatic capacitance reduction is indeed an intrinsic effect. We demonstrate the existence of a dielectric dead layer by calculating the dielectric profile across the interface and analyse its origin by extracting the ionic and electronic contributions to the electrostatic screening. We establish a correspondence between the dead layer and the hardening of the collective SrTiO₃ zone-centre polar modes, and determine the influence of the electrode by repeating our calculations for Pt/SrTiO₃/Pt capacitors. Our results provide practical guidelines for minimizing the deleterious effects of the dielectric dead layer in nanoscale devices.

The experimentally observed reduction in effective permittivity in nanoscale SrTiO₃ capacitors is usually expressed^{4–6} in terms of regions of low interfacial capacitance density, C_i , at the electrode–dielectric boundaries, acting in series with the nominal (classical) SrTiO₃ capacitance C_0 to give a lower overall capacitance C :

$$\frac{1}{C} = \frac{1}{C_i} + \frac{1}{C_0} + \frac{1}{C_i} \quad (1)$$

At present, two competing models for the interfacial layer exist. Arguments based on the finite electrostatic screening length in the metal support the idea of an intrinsic dead layer^{4,6,7}. Recent experiments, however, showed that, by using free-standing single-crystal lamellae instead of standard epitaxially grown films, bulk-like behaviour can be obtained even for structures as thin as 75 nm (ref. 8); the negligible observed shift in the Curie temperature indicates a remarkably high value for the interfacial capacitance. These results suggest that dead layers in thin films could arise from growth-induced defects, strains and grain boundaries⁹ rather than from the intrinsic properties of the ideal interface.

To complicate things further, the effect of the electrode material¹⁰ is even less clear. For example, the authors of ref. 11 pointed out that

the substantially better electronic screening provided by elemental metals should result in dead-layer-free capacitors. However, other authors⁶ suggested that the high ionic contribution to short-range polarizability provided by electrodes such as SrRuO₃ should yield improved screening; experiments on thin (Ba,Sr)TiO₃ films seem to support this argument^{5,10,12}. Finally, the reduction in permittivity has been shown to be correlated to a significant hardening of the 'soft' zone-centre optical phonon¹³, but a formal link between this effect and the interfacial dead layer has not been established.

First-principles calculations have long been used with great success to gain detailed and unbiased knowledge of the fundamental properties of metal–ceramic interfaces¹⁴. However, the lack of appropriate finite-field techniques has thwarted attempts to investigate the dielectric response of such systems. To remove this limitation, we recently introduced an important methodological advance in *ab initio* electronic structure theory that allows the introduction of a finite external field in the simulation of a periodic system with overall metallic character¹⁵. Our calculations in this work were carried out using the supercell method, with periodic heterostructures consisting of alternating seven-unit-cell slabs of insulating SrTiO₃ and metallic SrRuO₃ (or Pt). The in-plane cell parameter was fixed to the theoretical SrTiO₃ equilibrium lattice constant of 3.85 Å, to simulate epitaxial growth on a SrTiO₃ substrate. (The structural and dielectric properties of bulk SrTiO₃ calculated in this work are in excellent agreement with those reported in ref. 16, where they are extensively compared to the existing experimental data.) We relaxed the ionic positions and out-of-plane lattice constant (using an in-house implementation of density functional theory within the local density approximation and the projector-augmented wave method¹⁷) until both the out-of-plane stress and the maximum force on each ion were smaller than 1 meV Å^{−1}; the resulting energy offset between the Fermi level of the electrode and the SrTiO₃ valence band was 1.49 eV, very close to the value of 1.4 eV that was reported for the isostructural interface SrRuO₃/BaTiO₃ (ref. 18). We then calculated the symmetry-restricted force matrix¹⁹, the Born effective charges and the linear variation of the charge density upon small displacements of the ions. Using these, we extracted the capacitance, the spatially resolved permittivity profile and the full dynamical properties of the device in the linear (small field) limit. (The validity of the linear approach was confirmed by performing a direct lattice relaxation in a finite external bias of 0.05 eV, which yielded indistinguishable results.)

Our calculated static capacitance of the SrRuO₃/SrTiO₃ device (258 fF μm^{−2}) is very much reduced from the 'nominal' value $C_0 = (\epsilon_r \epsilon_0)/t = 1,600 \text{ fF } \mu\text{m}^{-2}$ (using our calculated bulk dielectric constant $\epsilon_r = 490$ and the thickness $t = 2.7 \text{ nm}$). There are no defects or strain fields in our dielectric film, so this is an intrinsic effect, which suggests that bulk values of capacitance are unattainable in SrRuO₃/SrTiO₃ nanocapacitors even if perfect metal–insulator interfaces could be prepared.

¹Materials Department, University of California, Santa Barbara, California 93106-5050, USA.

To analyse whether this dramatic reduction in the overall capacitance indeed results from a dielectric dead layer, we calculated the macroscopically averaged²⁰ local permittivity profile²¹, $\epsilon_r(x)$ (Fig. 1a). First, we note that, in the middle of either slab, $\frac{1}{\epsilon_r(x)}$ tends to the correct classical limit (respectively, 1/490 and 0). But the classical picture of a sharp transition between the bulk permittivity of SrTiO₃ and a diverging value in the metallic electrode clearly does not hold. Instead, at the interface there is a striking drop in the local permittivity, which appears as a strong peak in $\frac{1}{\epsilon_r(x)}$ roughly centred at the interfacial SrO layer [SrO(I)]. We identify this sub-polarized region as an ‘intrinsic dead layer’, which stems purely from the fundamental quantum-mechanical and electrostatic properties of the ideal metal–insulator interface. Equation (1) then yields an interface capacitance $C_i = 615 \text{ fF } \mu\text{m}^{-2}$ that compares well with the most recent experimental estimates¹⁰.

The deleterious effect of the dead layer on the static response of the device can be understood from the induced potentials shown in Fig. 1b. First we kept the atoms fixed to their zero-field positions (corresponding experimentally to probing the device at optical frequencies) and relaxed the electrons to their ground state within a fixed external bias $\Delta V = 27.8 \text{ mV}$. The resulting electrostatic potential is flat in the electrodes (the conduction electrons rearrange themselves to screen the macroscopic field inside the metal) and yields a uniform field of magnitude $E^\infty = 10.1 \text{ V } \mu\text{m}^{-1}$ in the middle of the dielectric; this situation closely matches the classical picture $E^{\text{classical}} = \Delta V/t \approx 10 \text{ V } \mu\text{m}^{-1}$. Once the atoms are allowed to relax, however, the field in the dielectric reduces to $E^0 = 1.64 \text{ mV } \mu\text{m}^{-1}$, accompanied by a potential drop in the interfacial low-permittivity layer. The difference $E_d = E^\infty - E^0$ is an electrostatic field that counteracts the collective relaxation of the atoms in an applied external bias; this phenomenon, which arises here because of the interfacial dead layer, is often indicated

as a depolarizing field^{22–24} in the context of ferroelectric thin-film capacitors.

The physics of the depolarizing field is usually discussed²⁵ in terms of the lattice-dynamical properties of the ‘soft’ zone-centre transverse optical phonon (TO1), which is responsible for the unusually high dielectric response of perovskite titanates. A zone-centre optical mode consists of a rigid relative translation of the ionic sublattices, which carries a macroscopically uniform dipole density in the infinite crystal. When the crystal is terminated by an interface, the dipolar field generates a surface charge density on either side of the slab, which is completely screened in the bulk by the cooperative periodic motion of the infinite lattice. A realistic electrode like SrRuO₃ is able to screen this perturbation only partially, and the residual unscreened surface charge generates the uniform depolarizing field²², E_d , that is linear in the magnitude ξ of the normal-mode displacement vector and is opposed to the polarization direction:

$$E_d = -k\xi \quad (2)$$

Because this restoring force provides a positive energy contribution ($U_d = 1/2k\xi^2$ in the harmonic limit), an unavoidable consequence of the depolarizing field is the ‘hardening’ of the soft mode of the film; in ferroelectric capacitors, this is fully consistent with the appearance of a critical thickness for ferroelectricity, which is determined by the balance between the strength of the lattice instability and the magnitude of the depolarizing effect, the latter dominating in thinner films.

To demonstrate this for our case, we calculate the phonon density of states of the system, weighted by the dipolar activity of each mode along the field axis (Fig. 2); this corresponds to the infrared absorption spectrum of the capacitor within short-circuit electrical boundary conditions. The three relevant zone-centre optical modes (TO1, TO2 and TO4) of bulk SrTiO₃ can be readily identified in the spectrum with the three most prominent peaks (A, B and C). As expected, their frequency is shifted to higher values, most dramatically in the case of the lowest-frequency soft mode (TO1).

In bulk SrTiO₃ the soft TO1 mode provides 99% of the contribution

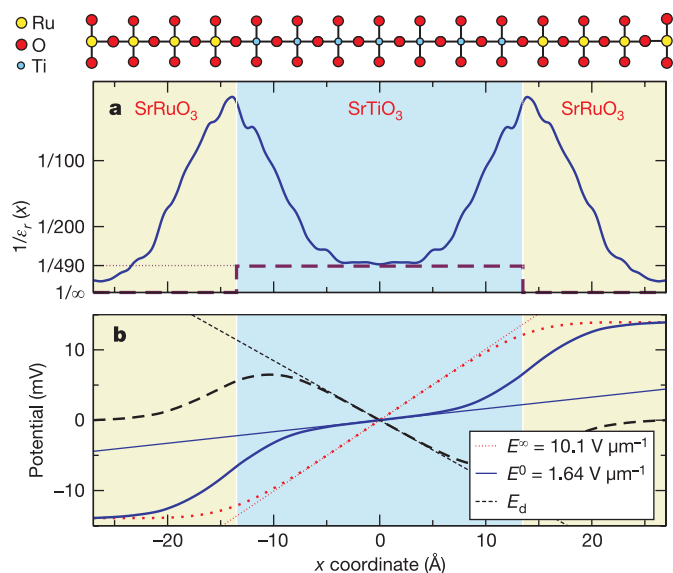


Figure 1 | Inverse permittivity profile and induced electrostatic potentials. **a**, Calculated inverse permittivity profile for the SrTiO₃/SrRuO₃ capacitor (blue solid curve) as compared to the classical picture (thick dashed curve) of a sharp transition between two ideal bulk materials. **b**, Induced electrostatic potentials for an external bias $\Delta V = 27.8 \text{ mV}$. First we relaxed the electrons while keeping the atoms frozen in their zero-field centrosymmetric positions (red dotted curve). Then we fully relaxed the structure within the external bias (blue solid curve). The difference between the two provides the depolarizing field E_d (black dashed curve). The magnitude of the electric field in the middle of the SrTiO₃ slab is given by the slope (thin lines). The potentials are flat in the electrodes, as expected from classical electrostatics (the conduction electrons rearrange themselves to screen the macroscopic field inside the metal).

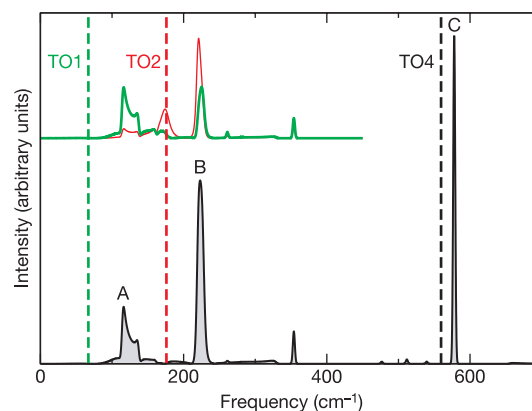


Figure 2 | Infrared absorption spectrum of the SrRuO₃/SrTiO₃ capacitor and decomposition into bulk polar modes. The symmetry-restricted (only polar modes along the field axis contribute to the capacitance) dynamical matrix of the system was diagonalized to obtain the normal modes, and the phonon density of states was weighted by the square of the mode effective charge¹⁶, which corresponds to the dipolar activity of the phonon. To optimize the frequency resolution, the force matrix was extended up to 201 layers of SrRuO₃ by using the bulk force constants and the curve was broadened by a gaussian width of 2 cm^{-1} . The frequencies of the most prominent peaks are $\omega_A = 116 \text{ cm}^{-1}$, $\omega_B = 225.5 \text{ cm}^{-1}$ and $\omega_C = 578 \text{ cm}^{-1}$. The calculated bulk phonon frequencies ($\omega_{\text{TO1}} = 67 \text{ cm}^{-1}$, $\omega_{\text{TO2}} = 176 \text{ cm}^{-1}$ and $\omega_{\text{TO4}} = 559.5 \text{ cm}^{-1}$) are shown as dashed vertical lines for comparison. The inset shows the projections of the phonon density of states onto the eigenvectors of the relevant zone-centre optical modes of bulk SrTiO₃: TO1 (thick green curve) and TO2 (thin red curve).

to the static permittivity, so we would expect the dead-layer-induced reduction in the capacitance density to be proportional to the square of the soft mode hardening¹³. Interestingly, this is not true in our system ($\omega_A^2/\omega_{TO1}^2 \approx 3$, $C_0/C \approx 6.2$). To investigate this discrepancy, we projected the eigenmodes of the heterostructure onto the eigenvectors of bulk SrTiO₃ zone-centre optical modes. Strikingly, while TO4 (not shown) remains isolated and preserves bulk-like properties, there is a strong interaction between TO1 and TO2 (Fig. 2 inset) which is induced by the presence of the interface. In particular, a significant part of the spectral weight of TO1 is transferred to the B peak; this further blueshift of part of the TO1-derived modes yields an overall capacitance significantly smaller than that suggested by the simple ratio of the frequencies. The mixing between TO1 and TO2 is also an unavoidable consequence of the dead-layer-induced depolarizing field. Taking into account all three infrared-active modes of SrTiO₃, the depolarizing field:

$$E_d = -k_1\xi_1 - k_2\xi_2 - k_3\xi_3 \quad (3)$$

couples with each of the three optical modes, so that interaction cross-terms naturally appear in the hamiltonian.

This analysis provides the long-sought reconciliation¹³ between the phenomenological theories of imperfect screening of the depolarizing field, the dead-layer effect and soft-mode hardening. The blueshift observed in the infrared absorption spectra is the reciprocal-space signature of the interfacial capacitance due to the intrinsic dead-layer effect; the anomalous behaviour that was detected in thinner films¹³ can be ascribed to the TO1–TO2 mixing, which is also a physical consequence of the dielectric dead layer at the interface.

Having identified the incomplete screening at the electrode as the main cause of both the interfacial dead layer and the soft-mode hardening, we can now gain further insight by separating the electronic and ionic contributions to the screening. It has been suggested⁶ that the high ionic polarizability of SrRuO₃ might compensate for the relatively inefficient electronic screening capabilities, and thus possibly eliminate any dielectric dead layer at the SrRuO₃/SrTiO₃ interface. To investigate this, we relaxed the structure with the SrRuO₃ atoms constrained to their zero-field positions; in Fig. 3 we compare the resulting permittivity profiles. The purely electronic screening of the polarization charges by SrRuO₃ is quite inefficient, and in the ‘frozen-SrO(I)’ set-up the interfacial capacitance is roughly twice as low as the fully relaxed value. The relaxation of SrO(I) provides the most important correction to the interfacial polarizability, while the rest of the SrRuO₃ slab contributes to a lesser

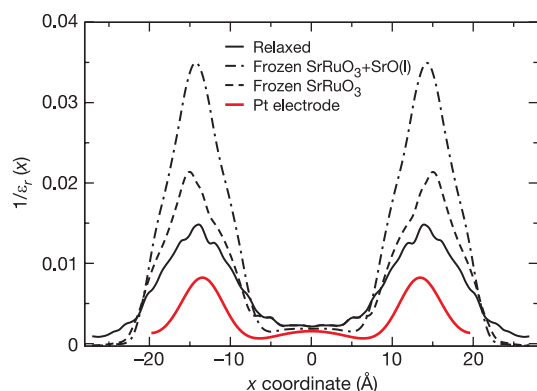


Figure 3 | Influence of the ionic contribution to the screening on the magnitude of the dead layer. Black lines show the inverse permittivity profiles obtained by freezing the ions in the SrRuO₃ electrodes. The fully relaxed result of Fig. 1 (solid curve) is compared to the frozen-electrode profiles, respectively allowing (dashed) or not allowing (dot-dashed) the interfacial SrO(I) layer to relax. The calculated profile for a Pt/SrTiO₃ capacitor (shown as a thick red curve) demonstrates the superior properties that can be achieved by using elemental metals as electrodes.

extent. Thus, the lattice polarization of SrRuO₃ significantly enhances the dielectric response of the capacitor, but is only able to reduce the dead layer by half, and not to remove it completely as was previously thought^{6,12,26}. We note that a recent report²⁷ of a twofold increase in the polarization of ferroelectric BaTiO₃/SrRuO₃ heterostructure upon ionic relaxation of the electrodes is in qualitative agreement with our results; the bound electron contribution to the screening in SrRuO₃ was recently also investigated experimentally²⁸.

It is clear from this analysis that excellent screening properties at the interface are crucial to reducing the dead layer; even one or two subpolarized atomic monolayers can seriously affect the properties of micrometre-thick devices. Our results for SrRuO₃/SrTiO₃ indicate that the ionic contribution to the screening, even with excellent structural matching between the electrode and the film, is unlikely ever to be sufficient. Therefore, electrodes with shorter electronic screening length, such as Pt or Au, should be the best candidates for reducing the dead layer. To check this conjecture, we repeated our calculations for a [Pt₂]₇SrO-[TiO₂-SrO]₆ capacitor, and used the same computational parameters as in the case of oxide electrodes. (To obtain a qualitatively correct band line-up at the interface, we used a geometry where the interfacial Sr and O atoms were aligned midway between the ‘atop’ and ‘hollow’ positions of the outermost Pt layer; the atomic positions were then fully relaxed along the field axis. In this way the SrTiO₃ valence band was 1.4 eV below the Fermi level of Pt.)

We find that the dead-layer effect is reduced by almost a factor of four (with an interfacial capacitance density $C_i = 2,350 \text{ fF } \mu\text{m}^{-2}$); the dramatic improvement is apparent in the calculated permittivity profile (Fig. 3). It is interesting to note that many experimental groups have reported an opposite trend¹⁰, that is, interfacial capacitances much smaller for Pt electrodes compared to SrRuO₃ ones. Our results strongly suggest that defects have a dominant role in determining the observed dielectric properties of (Ba,Sr)TiO₃/Pt devices, while in (Ba,Sr)TiO₃/SrRuO₃ the superior perovskite–perovskite epitaxial quality brings the measured C_i close to the fundamental limits calculated in this work. We note that recent calculations²⁹, showing enhanced polarization in ferroelectric PbTiO₃ films with Pt electrodes compared to metallic oxide ones, are in qualitative agreement with our conclusions. Finally, to demonstrate further that the higher interfacial capacitance is indeed the consequence of the improved electronic screening and not of specific details of this choice of materials and interface geometry, we repeated our calculations for an Au/BaZrO₃/Au capacitor, and obtained $C_i \approx 2,000 \text{ fF } \mu\text{m}^{-2}$, very similar to the Pt/SrTiO₃ case. This result confirms the generality of our arguments.

We note that our results are relevant not only for ultrathin devices; because the intrinsic interfacial capacitance acts in series with the central SrTiO₃ region and reduces the internal field over the whole film, it can significantly affect the properties of much thicker capacitors. For example, using equation (1) and our calculated parameters, a 75-nm-thick SrTiO₃/SrRuO₃ capacitor would have only 85% of the nominal capacitance, which is quite impressive considering that only one or two atomic monolayers at the interface are responsible for this reduction.

The challenge in the near future is then the production of atomically sharp interfaces between elemental metals and high- ϵ_r perovskites. Our hope is that the results presented here will stimulate further experimental work to improve the fabrication and characterization of such devices.

Received 24 March; accepted 31 July 2006.

1. Scott, J. F. High-dielectric constant thin films for dynamic random access memories (DRAM). *Annu. Rev. Mater. Sci.* **28**, 79–100 (1998).
2. Nagaraj, B. *et al.* (Ba,Sr)TiO₃ thin films with conducting perovskite electrodes for dynamic random access memory applications. *Appl. Phys. Lett.* **74**, 3194–3196 (1999).
3. Mead, C. A. Anomalous capacitance of thin dielectric structures. *Phys. Rev. Lett.* **6**, 545–546 (1961).

4. Zhou, C. & Newns, D. M. Intrinsic dead layer effect and the performance of ferroelectric thin film capacitors. *J. Appl. Phys.* **82**, 3081–3088 (1997).
5. Sinnamon, L. J., Bowman, R. M. & Gregg, J. M. Investigation of dead-layer thickness in $\text{SrRuO}_3/\text{Ba}_{0.5}\text{Sr}_{0.5}\text{TiO}_3/\text{Au}$ thin-film capacitors. *Appl. Phys. Lett.* **78**, 1724–1726 (2001).
6. Black, C. T. & Welser, J. J. Electric-field penetration into metals: consequences for high-dielectric constant capacitors. *IEEE Trans. Electron Devices* **46**, 776–780 (1999).
7. Natori, K., Otani, D. & Sano, N. Thickness dependence of the effective dielectric constant in a thin film capacitor. *Appl. Phys. Lett.* **73**, 632–634 (1998).
8. Saad, M. M. *et al.* Intrinsic dielectric response in ferroelectric nano-capacitors. *J. Phys. Condens. Matter* **16**, L451–L456 (2004).
9. Sinnamon, L. J., Saad, M. M., Bowman, R. M. & Gregg, J. M. Exploring grain size as a cause for “dead-layer” effects in thin film capacitors. *Appl. Phys. Lett.* **81**, 703–705 (2002).
10. Plonka, R., Dittmann, R., Pertsev, N. A., Vasco, E. & Waser, R. Impact of the top-electrode material on the permittivity of single-crystalline $\text{Ba}_{0.7}\text{Sr}_{0.3}\text{TiO}_3$ thin films. *Appl. Phys. Lett.* **86**, 202908 (2005).
11. Dawber, M. & Scott, J. F. Models of electrode-dielectric interfaces in ferroelectric thin-film devices. *Jpn. J. Appl. Phys.* **41**, 6848–6851 (2002).
12. Vendik, O. G., Zubko, S. P. & Ter-Martirosyan, L. T. Experimental evidence of the size effect in thin ferroelectric films. *Appl. Phys. Lett.* **73**, 37–39 (1998).
13. Sirenko, A. A. *et al.* Soft-mode hardening in SrTiO_3 thin films. *Nature* **404**, 373–376 (2000).
14. Finnis, M. W. The theory of metal-ceramic interfaces. *J. Phys. Condens. Matter* **8**, 5811–5836 (1996).
15. Stengel, M. & Spaldin, N. A. *Ab-initio* theory of metal-insulator interfaces in a finite electric field. Preprint at (<http://www.arXiv.org/cond-mat/0511042>) (2005).
16. Antons, A., Neaton, J. B., Rabe, K. M. & Vanderbilt, D. Tunability of the dielectric response of epitaxially strained SrTiO_3 from first-principles. *Phys. Rev. B* **71**, 024102 (2005).
17. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
18. Junquera, J., Zimmer, M., Ordejón, P. & Ghosez, Ph. First-principles calculation of the band offset at $\text{BaO}/\text{BaTiO}_3$ and $\text{SrO}/\text{SrTiO}_3$ interfaces. *Phys. Rev. B* **67**, 155325 (2003).
19. Baroni, S., de Gironcoli, S. & Corso, A. D. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73**, 515–562 (2001).
20. Baldereschi, A., Baroni, S. & Resta, R. Band offsets in lattice-matched heterojunctions: a model and first-principles calculations for GaAs/AlAs . *Phys. Rev. B* **61**, 734–737 (1988).
21. Giustino, F. & Pasquarello, A. Theory of atomic-scale dielectric permittivity at insulator interfaces. *Phys. Rev. B* **71**, 144104 (2005).
22. Junquera, J. & Ghosez, Ph. Critical thickness for ferroelectricity in perovskite ultrathin films. *Nature* **422**, 506–509 (2003).
23. Spaldin, N. A. Fundamental size limits in ferroelectricity. *Science* **304**, 1606–1607 (2004).
24. Fong, D. D. *et al.* Ferroelectricity in ultrathin perovskite films. *Science* **304**, 1650–1653 (2004).
25. Ghosez, Ph. & Junquera, J. in *Handbook of Theoretical and Computational Nanotechnology* (eds Rieth, M. & Schommers, W.) (American Scientific Publisher, Stevenson Ranch, California, 2006).
26. Hwang, C. S. Thickness-dependent dielectric constants of $(\text{Ba},\text{Sr})\text{TiO}_3$ thin films with Pt or conducting oxide electrodes. *J. Appl. Phys.* **92**, 432–437 (2002).
27. Gerra, G., Tagantsev, A. K., Setter, N. & Parlinski, K. Ionic polarizability of conductive metal oxides and critical thickness for ferroelectricity in BaTiO_3 . *Phys. Rev. Lett.* **96**, 107603 (2006).
28. Kim, D. J. *et al.* Polarization relaxation induced by a depolarization field in ultrathin ferroelectric BaTiO_3 capacitors. *Phys. Rev. Lett.* **95**, 237602 (2005).
29. Sai, N., Kolpak, A. M. & Rappe, A. M. Ferroelectricity in ultrathin perovskite films. *Phys. Rev. B* **72**, 020101(R) (2005).

Acknowledgements We thank A. De Vita, Ph. Ghosez, J. Junquera, R. Ramesh, R. Seshadri, J. F. Scott and S. Stemmer for discussions. This work was supported by the National Science Foundation's Division of Materials Research, through the Condensed Matter and Materials Theory Program; computations were performed on the DataStar machine at the San Diego Supercomputer Center.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (stengel@mrl.ucsb.edu).

Bistability of atmospheric oxygen and the Great Oxidation

Colin Goldblatt^{1,2}, Timothy M. Lenton¹ & Andrew J. Watson¹

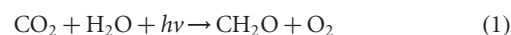
The history of the Earth has been characterized by a series of major transitions separated by long periods of relative stability¹. The largest chemical transition was the 'Great Oxidation', approximately 2.4 billion years ago, when atmospheric oxygen concentrations rose from less than 10^{-5} of the present atmospheric level (PAL) to more than 0.01 PAL, and possibly² to more than 0.1 PAL. This transition took place long after oxygenic photosynthesis is thought to have evolved^{3–5}, but the causes of this delay and of the Great Oxidation itself remain uncertain^{6–11}. Here we show that the origin of oxygenic photosynthesis gave rise to two simultaneously stable steady states for atmospheric oxygen. The existence of a low-oxygen (less than 10^{-5} PAL) steady state explains how a reducing atmosphere persisted for at least 300 million years after the onset of oxygenic photosynthesis. The Great Oxidation can be understood as a switch to the high-oxygen (more than 5×10^{-3} PAL) steady state. The bistability arises because ultraviolet shielding of the troposphere by ozone becomes effective once oxygen levels exceed 10^{-5} PAL, causing a nonlinear increase in the lifetime of atmospheric oxygen. Our results indicate that the existence of oxygenic photosynthesis is not a sufficient condition for either an oxygen-rich atmosphere or the presence of an ozone layer, which has implications for detecting life on other planets using atmospheric analysis^{12,13} and for the evolution of multicellular life.

Before the origin of life, photochemical models¹⁴ predict oxygen levels to have been $O_2 < 10^{-12}$ PAL. The recent discovery of mass independent fractionation (MIF) of sulphur isotopes¹⁵ before 2.45 billion years ago (Gyr ago), constrains¹⁶ O_2 to $< 10^{-5}$ PAL. The MIF signal is absent in samples 2.33 Gyr old and younger¹⁷, indicating that the Great Oxidation occurred approximately 2.4 Gyr ago. Geological and biological constraints indicate oxygen levels consistently > 0.01 PAL after the Great Oxidation, and possibly > 0.1 PAL. Compilations of MIF data and all geological constraints on palaeo-oxygen are included as Supplementary Information. The Great Oxidation would be most easily understood as the immediate consequence of the origin of oxygenic photosynthesis¹⁸, but evidence from biomarkers^{3–5} suggests this innovation had occurred at least 300 Myr earlier, by 2.7 Gyr ago. Recent hypotheses for the cause are: an increase in the oxidation state of mantle outgassing^{6,8}, but this appears to contradict geological constraints^{9,19}; or a progressive oxidation of the crust, leading to a decreased flux of metamorphic reductants^{10,11}, but the mechanism of the required oxidation is unclear⁹. These hypotheses assume that the Great Oxidation was triggered when the oxygen source exceeded the input of volcanic and metamorphic reductants^{6–9,11}.

Our conceptual model of the global reduction–oxidation (redox) system (Fig. 1) explains the stability of oxygen levels each side of the Great Oxidation, the delay from the origin of oxygenic photosynthesis and the apparent rapidity and irreversibility of the transition. The rationale for the model is given here and it is described fully in

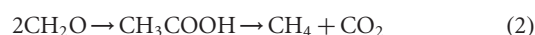
the Methods and Supplementary Information.

Organisms play a pivotal role in the global redox system, generating fluxes of both oxidized and reduced gases to the atmosphere. Oxygenic photosynthesis is summarized thus:

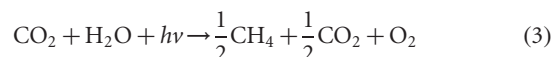


where $h\nu$ indicates photosynthetically active radiation.

Today, around half of the carbon fixed is consumed internally by photosynthetic organisms in aerobic respiration, the reverse of equation (1). The remainder, the net primary productivity, is available to other organisms. The principal metabolic path for this carbon today is heterotrophic aerobic respiration, but this is inhibited below the Pasteur point, $O_2 \approx 0.01$ PAL. The metabolic pathway effective in anoxic conditions is fermentation followed by acetogenic methanogenesis:



Oxygenic photosynthesis followed by fermentation and methanogenesis is summarized:



This is likely to have been the major metabolic path before the Great Oxidation. If there is sufficient ambient oxygen ($O_2 > 2 \times 10^{-4}$ PAL), methane-oxidizing bacteria (methanotrophs) will consume

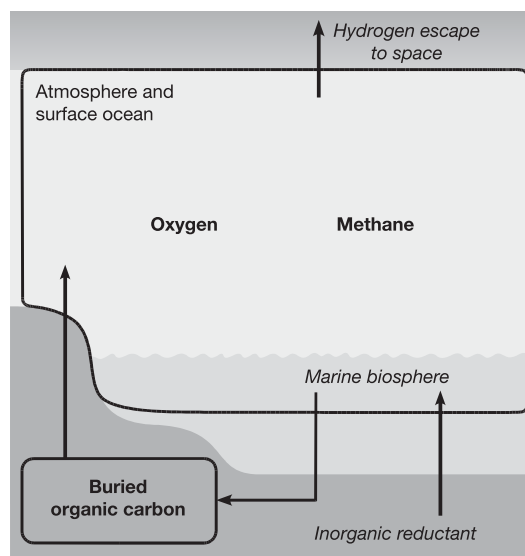


Figure 1 | Model schematic. We resolve methane and oxygen abundances in the atmosphere and surface ocean and a buried organic carbon in the crust.

¹School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK. ²Centre for Ecology and Hydrology, Edinburgh Research Station, Bush Estate, Penicuik, Midlothian, EH26 0QB, UK.

some of the methane and oxygen produced:



but the remainder will be a flux of both methane and oxygen to the atmosphere in a 1:2 stoichiometric ratio. The fraction of the net primary productivity which has this fate, Ω_{O_2} , is higher in a low-oxygen atmosphere.

We describe the overall redox balance of the atmosphere–ocean system with a net input of reducing material represented by an input r of ferrous iron, burial and weathering of organic carbon and an oxidation due to loss of hydrogen to space sM , where s is a limiting flux constant¹⁰ and M is the amount of methane in the atmosphere (see Supplementary Information).

Once oxygenic photosynthesis is well established with assumed net primary productivity N , the oxygen flux from the biosphere is $\Omega_{\text{O}_2}N$, representing the oxygen source from oxygenic photosynthesis, with the oxygen used in heterotrophic respiration and by methanotrophs subtracted. The corresponding methane flux is $\frac{1}{2}\Omega_{\text{O}_2}N$, from methanogenic decomposition of organic matter with consumption by methanotrophs subtracted. Thus the early Earth's biosphere, like today's, promoted an atmosphere in strong thermodynamic disequilibrium^{12,13}. The dominant process in restoring equilibrium is atmospheric methane oxidation. This can be summarized with the stoichiometry of equation (4), but takes place as a series of reactions, some of which are photochemically mediated. We empirically parameterize this flux as a function of methane and oxygen abundances, fitting it to the results of detailed photochemical models^{16,20,21} that resolve these reactions, obtaining oxidation rate $\Psi_{\text{O}_2}M^{0.7}$, where Ψ_{O_2} is a polynomial function of oxygen concentration (see Supplementary Information). An important property of the atmospheric chemistry embodied in Ψ_{O_2} is that, once oxygen has increased to a certain level, an ozone layer forms, shielding the troposphere from ultraviolet radiation. This dramatically decreases the rate of photolysis of water vapour in the troposphere, reducing hydroxyl radical availability, and thus suppressing the rate of methane oxidation^{16,22}.

Proxies indicate that either side of the Great Oxidation oxygen concentrations were relatively stable for 10^8 – 10^9 years, so these should correspond to steady-state solutions of our model. Steady-state methane balances net input r and output sM of reductant from the atmosphere–ocean system, giving:

$$M = \frac{r}{s} \quad (5)$$

The steady state for oxygen is dominated by the balance of production by the biosphere and loss of oxidizing methane; by substituting $M = r/s$ oxygen abundance is found as the solutions of:

$$\Omega_{\text{O}_2}N - \Psi_{\text{O}_2}\left(\frac{r}{s}\right)^{0.7} = 0 \quad (6)$$

There are two sets of stable steady-state solutions for oxygen in the presence of oxygenic photosynthesis (Fig. 2); high oxygen ($\text{O}_2 > 5 \times 10^{-3}$ PAL) and low oxygen ($\text{O}_2 < 10^{-5}$ PAL). Where these overlap, they are separated by an unstable steady-state solution. The cause of the bistability is a strong nonlinear positive feedback on oxygen levels. Once oxygen exceeds 2×10^{-5} PAL an ozone layer starts to form, decreasing the rate of methane oxidation. This reduced oxygen sink causes oxygen levels to increase, promoting the formation of the ozone layer. Oxygen continues to rise until the photochemical loss again balances the biospheric source. The level of the high-oxygen stable steady state is decreased by negative feedback within the biosphere; as oxygen increases, methanotrophy and heterotrophic respiration become effective at thresholds of 2×10^{-4} PAL and 0.01 PAL respectively, decreasing the fluxes of methane and oxygen to the atmosphere.

The low-oxygen solution is consistent with the MIF constraint of $\text{O}_2 < 10^{-5}$ PAL before the Great Oxidation. It is maintained in the presence of oxygenic photosynthesis, explaining the time lag between the origin of oxygenic photosynthesis and the Great Oxidation. The

high-oxygen solution is consistent with proxy constraints following the Great Oxidation (Supplementary Fig. 1).

We hypothesize that the Great Oxidation corresponded to a switch between stable steady states when oxygen exceeded 10^{-5} PAL and an ozone layer developed. This contrasts with previous assumptions^{6–9,11} that it occurred when the oxygen source from carbon burial exceeded the input of volcanic and metamorphic reductants. The system was probably in the bistable region shortly before the Great Oxidation (Fig. 2). Increasing N , or decreasing r , would mean that only the high-oxygen solution could exist, forcing the switch. A perturbation to the carbon cycle could also cause the transition, with the system remaining in the bistable region. Two plausible scenarios are outlined below.

Contemporaneous with the Great Oxidation, there is a marked decrease in banded iron formations (BIFs) in the sedimentary record approximately 2.4 Gyr ago (ref. 23), possibly caused by the cessation of a mantle superplume event⁷. Taking BIF deposition as a proxy for r implies a significant decrease in r (Fig. 2). With present productivity, the enhanced reductant input during BIF deposition² would correspond to oxygen in the bistable region ($\text{O}_2 \sim 10^{-6}$ PAL for the low-oxygen solution), whereas the present input of reduced iron to the ocean, with present productivity, corresponds to the monostable region of the high-oxygen solution ($\text{O}_2 \sim 10^{-1}$ PAL). Hence, decreasing r can trigger the Great Oxidation in our model (Fig. 3a). To maintain redox balance, the methane concentration of the atmosphere adjusts such that the loss of hydrogen to space balances the net input of reductant to the surface system. This adjustment takes around 50 million years during which there is a net oxidation as the

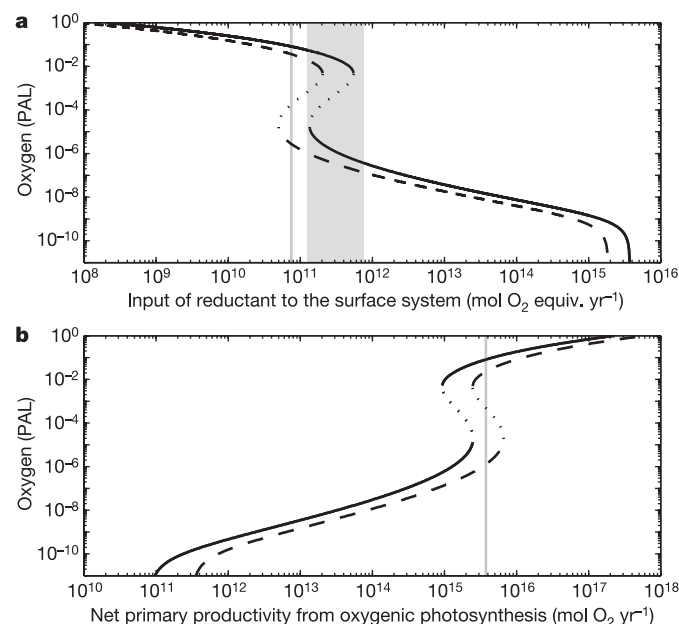


Figure 2 | Steady-state solutions for oxygen. Stable steady-state solutions are shown with solid and dashed lines, and unstable steady states with dotted lines. These are calculated with the full equation given in the Methods section; the simplified version given in the text is equivalent when oxygen exceeds 1×10^9 PAL. **a**, The solid line indicates modern marine net primary productivity ($N = 3.75 \times 10^{15}$ mol O_2 yr^{-1}) and the dashed line indicates half of the present N value for comparison. The shaded area shows constraints on iron deposition in BIFs before the Great Oxidation, which is taken as a proxy for reductant input. The minimum rate of iron deposition in the Hamersley BIF (2.69–2.44 Gyr ago; ref. 23) was 5×10^{11} mol Fe yr^{-1} and an upper bound for contemporaneous global deposition was 3×10^{12} mol Fe yr^{-1} (ref. 2), hence $1.25 \times 10^{11} < r < 7.5 \times 10^{11}$ mol O_2 equiv. yr^{-1} . The vertical grey line shows the present hydrothermal input of Fe to the ocean: 3×10^{11} mol Fe yr^{-1} (ref. 2), so $r = 7.5 \times 10^{10}$ mol O_2 equiv. yr^{-1} . **b**, The solid line indicates the present input of reduced iron to ocean, the dashed line indicates $r = 3 \times 10^{10}$ mol O_2 equiv. yr^{-1} , a possible value before the Great Oxidation. Modern marine N is shown as a vertical grey line.

system moves to its new steady state (Fig. 3b).

An alternative scenario is a small increase in N , and hence in organic carbon burial, with no change in r (Fig. 3c). As the size of the organic carbon reservoir adjusts, there is an oxygen source from excess organic carbon burial over weathering. Oxygen rises sufficiently to start ozone formation which causes the switch to high oxygen. Methane concentration, and hence oxidation from hydrogen escape, decreases transiently as the system adjusts. As a new steady state is achieved, organic carbon burial is balanced by weathering, atmospheric methane recovers and the net input of reductant is again balanced by hydrogen escape (Fig. 3d). The 3% increase in organic carbon burial assumed in this simulation would give a 0.23‰ increase in $\delta^{13}\text{C}$ carbonate (see Methods), which is less than the noise in that record²⁴. Thus the Great Oxidation could have been triggered by a carbon cycle perturbation too small to be seen in the carbon isotope record.

There were probably fluctuations in the organic carbon cycle of this magnitude between the origin of oxygenic photosynthesis and the Great Oxidation. The low-oxygen solution in the bistable region resists perturbations poorly, so the transition to the high-oxygen solution would be likely. This suggests that the system was in the monostable low-oxygen region for most of this time and that the Great

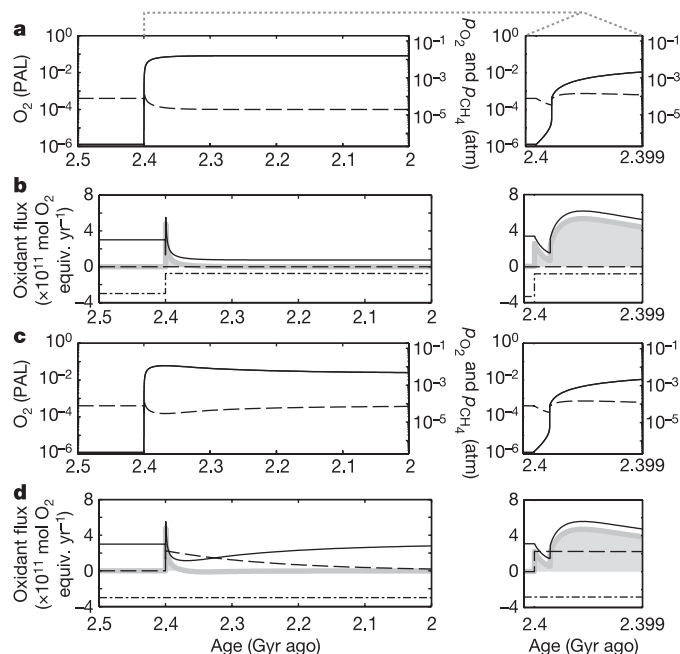


Figure 3 | Transient response to step changes at 2.4 Gyr ago representing possible triggers of the Great Oxidation. **a**, Response of oxygen (solid line, either scale) and methane (dashed line, right-hand scale) to a decrease in reductant input r from 3×10^{11} to 7.5×10^{10} mol O_2 equiv. yr^{-1} , inferred from decreased BIF deposition^{2,23} with present $N = 3.75 \times 10^{15}$ mol $\text{O}_2 \text{ yr}^{-1}$. There is no significant change in the amount of buried organic carbon. **b**, Redox budget (boundary fluxes) of the surface system corresponding to **a**. Hydrogen escape is shown as a solid black line, the net change in the size of the organic carbon reservoir (burial minus weathering) as a dashed line and reductant input as a dot-dashed line. The grey line is the summation of these terms, the net oxidation rate of the surface system. The integral of this (shaded) is the total oxidation of the system, 3×10^{18} mol, which corresponds to the oxygen content of the atmosphere at the end of the model run. **c**, Response of oxygen and methane to a 3% increase in N from 3.6375×10^{15} to 3.75×10^{15} mol $\text{O}_2 \text{ yr}^{-1}$ with constant $r = 3 \times 10^{11}$ mol O_2 equiv. yr^{-1} . In response, the amount of buried organic carbon increases by 3% over ~ 500 million years after the Great Oxidation. As in **a**, the solid line indicates the oxygen response, and the dashed line the methane response. **d**, Redox budget of the surface system corresponding to **c**. The total oxidation is 9×10^{17} mol. The grey line and shading are as in **b**. The right panels expand the transitions shown in the left panels.

Oxidation was triggered by a decrease in reductant moving the system to the bistable region, followed by a small carbon cycle perturbation.

The bistability in oxygen describes a typical hysteresis loop (Fig. 2). If the Great Oxidation was triggered by a small decrease in r , causing a shift from the bistable lower branch to the monostable upper branch, a much larger increase in r would have been needed to switch the system back to the low-oxygen state. The transition from low to high oxygen (Fig. 3) would take 150,000 years if forced only by decreasing r , or 50,000 years with a carbon cycle perturbation, but the reverse would take 3–7 million years, owing to the larger reservoir of oxygen and smaller rate constant for methane oxidation, making high oxygen more resistant to perturbation. This may explain why oxygen never returned to Archean values after the Great Oxidation².

Our model suggests an earlier rise of oxygen from prebiotic levels of $\sim 10^{-12}$ PAL to the low-oxygen solution (Fig. 2). This occurred when the oxygen source from the biosphere, equal to N in anoxic conditions ($\Omega_{\text{O}_2} \rightarrow 1$), exceeded r . This probably occurred directly after the origin of oxygenic photosynthesis.

It has been argued that the Great Oxidation caused a large decline in atmospheric methane concentration, leading to extreme glaciation in the Palaeoproterozoic^{25,26}. In transient runs of our model (see the right panels of Fig. 3), there is a rapid decrease in methane concentration by a factor of 2–6 while oxygen rises to the $>10^{-5}$ PAL threshold. The associated temperature drop is estimated to be 4–9 °C (see Supplementary Information). This may well have been sufficient to trigger extreme glaciation.

METHODS

Model structure. The model consists of a box representing the atmosphere and ocean mixed layer, in which the number of moles of O_2 (O) and CH_4 (M) are calculated, and a reservoir of organic carbon in the crust (C). Atmospheric gases are assumed to be dissolved in the mixed layer to saturation. The molar concentration of dissolved oxygen $[O] = (\alpha/\mu)O$ where $\mu = 1.773 \times 10^{20}$ mol is the number of moles in the present atmosphere and solubility is taken as $\alpha = 0.0013$ M.

Reductant input. The net input of inorganic reductant to the system is set as a boundary condition of r mol O_2 equiv. yr^{-1} , and is represented chemically by ferrous iron. It can include weathering of inorganic reductant in the crust and direct input from the mantle. The deep ocean is assumed to be anoxic, so any ferrous iron input here is taken to be transferred into the surface ocean. Anoxic photosynthesis in the surface ocean uses this ferrous iron as an electron donor:

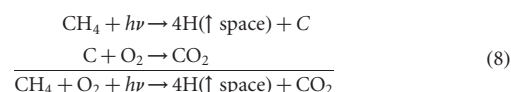


and the resulting $\text{Fe}(\text{OH})_3$ is precipitated as a BIF²⁷. Hence reducing power is transferred to organic carbon.

Marine biosphere. Oxygenic photosynthesis is represented by an assumed net primary productivity of N mol $\text{O}_2 \text{ yr}^{-1}$, so the total organic carbon produced is $(N+r)$. The fraction of this consumed by heterotrophic respirers is $\gamma = O/(d_\gamma + O)$, where $d_\gamma = 1.36 \times 10^{19}$ mol, corresponding to the inhibition of aerobic respiration at the Pasteur point, $\text{O}_2 \approx 0.01$ PAL. The remaining organic matter is decomposed by fermenters and acetogenic methanogens. The fraction of the methane produced which is consumed by methanotrophs is $\delta = O/(d_\delta + O)$, where $d_\delta = 2.73 \times 10^{17}$ mol, equivalent to $[O] = 2 \mu\text{M}$ (ref. 28). Thus the fraction of the oxygen and methane produced that reaches the atmosphere is given $\Omega_{\text{O}_2} = (1-\gamma)(1-\delta)$.

Atmospheric chemistry. Fitting the methane oxidation rate to the results of detailed photochemical models, we obtain $\Psi M^{0.7}$, where M is the number of moles of CH_4 , $\Psi = 10^{a_1\psi^4 + a_2\psi^3 + a_3\psi^2 + a_4\psi + a_5}$, $\psi = \log O$, $a_1 = 0.0030$, $a_2 = -0.1655$, $a_3 = 3.2305$, $a_4 = -25.8343$ and $a_5 = 71.5398$, valid for $10^8 < O < 10^{20}$ mol. This is derived in the Supplementary Information.

Hydrogen escape. Hydrogen escape to space is assumed to be diffusion limited, and so directly proportional to the total hydrogen mixing ratio in the upper atmosphere. With methane as the sole hydrogen-bearing species in our model, hydrogen loss is proportional to methane abundance, with rate sM where $s = 2.03 \times 10^{-5} \text{ yr}^{-1}$ (ref. 10). Stoichiometrically, this is represented as¹⁰:



This sink of methane and oxygen has a stoichiometric ratio of 1:1, but they are

produced in a ratio of 1:2, so hydrogen loss is a net source of oxidant. Water vapour is not a strong source of hydrogen for loss to space because little is able to pass the tropopause.

Crustal carbon cycle. Organic carbon burial is $\beta(N+r)$, a fraction of the organic carbon production. The rate of organic carbon weathering is wC , where the bulk weathering rate, $w = 6 \times 10^{-9} \text{ yr}^{-1}$, is set by analogy to present conditions and C is crustal organic carbon. At high oxygen levels, weathered material is oxidized directly. At low oxygen levels, oxidative weathering is ineffective and we assume that exposed organic carbon is consumed by organisms that transfer the reducing power to methane²⁹, which is subsequently oxidized in the atmosphere.

Model equations. The full algebraic derivation of our model is given in the Supplementary Information. The model equations are:

$$\frac{dM}{dt} = \frac{1}{2}\Omega_{\text{O}_2}(N+r) - sM - \frac{1}{2}\Psi_{\text{O}_2}M^{0.7} - \frac{1}{2}\Omega_{\text{O}_2}(\beta(N+r) - wC) \quad (9)$$

$$\frac{dO}{dt} = \Omega_{\text{O}_2}N - (1 - \Omega_{\text{O}_2})r - sM - \Psi_{\text{O}_2}M^{0.7} + (1 - \Omega_{\text{O}_2})(\beta(N+r) - wC) \quad (10)$$

$$\frac{dC}{dt} = \beta(N+r) - wC \quad (11)$$

Steady-state solutions are found by setting equations (9)–(11) to zero. Rearranging gives $M = r/s$ and $C = \beta(N+r)/w$. Oxygen is found as the roots of $\Omega_{\text{O}_2}N + (\Omega_{\text{O}_2} - 2)r - \Psi_{\text{O}_2}(r/s)^{0.7} = 0$. When $N \gg r$ these simplify to $C = \beta N/w$ and $\Omega_{\text{O}_2}N - \Psi_{\text{O}_2}(r/s)^{0.7} = 0$ (see Supplementary Information).

Carbon isotope calculations. The standard interpretation of the carbonate carbon isotope record is $f = (\delta_{\text{carb}} - \delta_i)/(\delta_{\text{carb}} - \delta_{\text{org}})$ where f is the fraction of volcanic CO_2 buried as organic carbon and δ_{carb} , δ_i and δ_{org} are the $\delta^{13}\text{C}$ signals of carbonate carbon, volcanic CO_2 and organic carbon respectively. Typical values³⁰ are $\delta_{\text{carb}} = 0\text{‰}$, $\delta_i = -6\text{‰}$ and $\delta_{\text{org}} = -30\text{‰}$, giving $f = 0.2$. Our 3% increase in organic carbon burial (Fig. 3d) would correspond to $f = 0.206$ and hence $\delta_{\text{carb}} = 0.23\text{‰}$.

Received 30 March; accepted 14 August 2006.

- Lenton, T., Caldeira, K. & Szathmáry, E. in *Earth Systems Analysis Sustainability* (eds Schellnhuber, H., Crutzen, P., Clark, W., Claussen, M. & Held, H.) 29–52 (Dahlem Workshop Reports, Vol. 91, MIT Press, Cambridge, Massachusetts, 2004).
- Holland, H. D. The oxygenation of the atmosphere and oceans. *Phil. Trans. Soc. B* **361**, 903–916 (2006).
- Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. Archean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033–1036 (1999).
- Summons, R. E., Jahnke, L. L., Hope, J. M. & Logan, G. A. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400**, 554–557 (1999).
- Brocks, J. J., Buick, R., Summons, R. E. & Logan, G. A. A reconstruction of Archean biological diversity based on molecular fossils from the 2.78 to 2.45 billion-year-old Mount Bruce Supergroup, Hamersley Basin, Western Australia. *Geochim. Cosmochim. Acta* **67**, 4321–4335 (2003).
- Holland, H. D. Volcanic gases, black smokers, and the great oxidation event. *Geochim. Cosmochim. Acta* **66**, 3811–3826 (2002).
- Barley, M. E., Bekker, A. & Krapež, B. Late Archean to early Paleoproterozoic global tectonics, environmental change and the rise of atmospheric oxygen. *Earth Planet. Sci. Lett.* **238**, 156–171 (2005).
- Kump, L. R., Kasting, J. F. & Barley, M. E. Rise of atmospheric oxygen and the “upside-down” Archean mantle. *Geochim. Geophys. Geosyst.* **2**, doi:10.1029/2000GC000114 (2001).
- Kasting, J. F. The rise of atmospheric oxygen. *Science* **293**, 819–820 (2001).
- Catling, D., Zahnle, K. & McKay, C. Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. *Science* **293**, 839–843 (2001).
- Catling, D. C. & Claire, M. W. How Earth’s atmosphere evolved to an oxic state: a status report. *Earth Planet. Sci. Lett.* **237**, 1–20 (2005).
- Lovelock, J. A physical basis for life detection experiments. *Nature* **207**, 568–570 (1965).
- Hitchcock, D. & Lovelock, J. Life detection by atmospheric analysis. *Icarus* **7**, 149–159 (1967).
- Kasting, J. F., Liu, S. C. & Donahue, T. M. Oxygen levels in the prebiological atmosphere. *J. Geophys. Res.* **84**, 3097–3107 (1979).
- Farquhar, J., Bao, H. & Thieme, M. Atmospheric influence of Earth’s earliest sulfur cycle. *Science* **289**, 757–758 (2000).
- Pavlov, A. & Kasting, J. Mass-independent fractionation of sulfur isotopes in Archean sediments: strong evidence for an anoxic Archean atmosphere. *Astrobiology* **2**, 27–41 (2002).
- Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).
- Kopp, R. E., Kirschvink, J. L., Hilburn, I. A. & Nash, C. Z. The paleoproterozoic snowball Earth: A climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc. Natl Acad. Sci. USA* **32**, 11131–11136 (2005).
- Li, Z.-X. A. & Lee, C.-T. A. The constancy of upper mantle $f\text{O}_2$ through time inferred from V/Sc ratios in basalts. *Earth Planet. Sci. Lett.* **228**, 483–493 (2004).
- Pavlov, A., Brown, L. & Kasting, J. UV shielding of NH_3 and O_2 by organic hazes in the Archean atmosphere. *J. Geophys. Res.* **106**, 23267–23287 (2001).
- Pavlov, A., Hurtgen, M., Kasting, J. & Arthur, M. Methane-rich Proterozoic atmosphere? *Geology* **31**, 87–90 (2003).
- Kasting, J. & Donahue, T. The evolution of the atmospheric ozone. *J. Geophys. Res.* **85**, 3255–3263 (1980).
- Isley, A. E. & Abbott, D. H. Plume-related mafic volcanism and the deposition of banded iron formation. *J. Geophys. Res.* **104**, 15461–15477 (1999).
- Shields, G. & Veizer, J. Precambrian marine carbonate isotope database: version 1.1. *Geochim. Geophys. Geosyst.* **3**, doi:10.1029/2001GC000266 (2002).
- Pavlov, A. A., Kasting, J. F., Brown, L. L., Rages, K. A. & Freedman, R. Greenhouse warming by CH_4 in the atmosphere of early Earth. *J. Geophys. Res.* **105**, 11981–11990 (2000).
- Kasting, J. F. Methane and climate during the precambrian era. *Precambrian Res.* **137**, 119–129 (2005).
- Kharecha, P., Kasting, J. F. & Siefert, J. L. A coupled atmosphere-ecosystem model of the early Archean Earth. *Geobiology* **3**, 53–76 (2005).
- Ren, T., Amaral, J. A. & Knowles, R. The response of methane consumption by pure cultures of methanotropic bacteria to oxygen. *Can. J. Microbiol.* **43**, 925–928 (1997).
- Holland, H. D. Discussion of the article by A. C. Lasaga and H. Ohmoto on “The oxygen geochemical cycle: dynamics and stability”, *Geochim. Cosmochim. Acta* **66**, 361–381, 2002. *Geochim. Cosmochim. Acta* **67**, 787–789 (2003).
- Hayes, J. M. & Waldbauer, J. R. The carbon cycle and associated redox processed through time. *Phil. Trans. R. Soc. B* **361**, 931–950 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements C.G. is funded by the Natural Environment Research Council. T.M.L. is supported by a Leverhulme Prize. We thank J. Kasting and L. Kump for comments on the manuscript. C.G. thanks J. Meiss for teaching dynamical systems.

Author Contributions T.M.L. and A.J.W. suggested the study. C.G. wrote and analysed the model with supervision from T.M.L. and A.J.W., and led writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare that they have no competing financial interests. Correspondence and requests for materials should be addressed to C.G. (c.goldblatt@uea.ac.uk).

Long-period astronomical forcing of mammal turnover

Jan A. van Dam¹, Hayfaa Abdul Aziz^{1†}, M. Ángeles Álvarez Sierra², Frederik J. Hilgen¹, Lars W. van den Hoek Ostende³, Lucas J. Lourens¹, Pierre Mein⁴, Albert J. van der Meulen¹ & Pablo Pelaez-Campomanes⁵

Mammals are among the fastest-radiating groups, being characterized by a mean species lifespan of the order of 2.5 million years (Myr)^{1,2}. The basis for this characteristic timescale of origination, extinction and turnover is not well understood. Various studies have invoked climate change to explain mammalian species turnover^{3,4}, but other studies have either challenged or only partly confirmed the climate–turnover hypothesis^{5–7}. Here we use an exceptionally long (24.5–2.5 Myr ago), dense, and well-dated terrestrial record of rodent lineages from central Spain, and show the existence of turnover cycles with periods of 2.4–2.5 and 1.0 Myr. We link these cycles to low-frequency modulations of Milankovitch oscillations⁸, and show that pulses of turnover occur at minima of the 2.37-Myr eccentricity cycle and nodes of the 1.2-Myr obliquity cycle. Because obliquity nodes and eccentricity minima are associated with ice sheet expansion and cooling and affect regional precipitation, we infer that long-period astronomical climate forcing is a major determinant of species turnover in small mammals and probably other groups as well.

Changes in organisms occur on a variety of temporal scales. Four basic temporal scales (tiers) have been recognized: the ecological timescale; the Milankovitch timescale of precession (the wobbling of the Earth's axis, ~21 kyr periodicity), obliquity (tilt of the Earth's axis, 41 kyr periodicity) and eccentricity (the orbit around the Sun, ~100 and ~400 kyr periodicity); the million-year timescale of species extinctions and originations; and the ultra-long timescale of mass extinctions and major taxonomic replacements⁹. Although the main processes controlling the first tier (climate change and competition), the second tier (climate-forced distributional variation)^{3,9,10} and the fourth tier (catastrophic perturbations of the Earth's biosphere) are reasonably well defined, the mechanisms underlying third-tier processes are not well understood. Mammals have featured importantly in discussions on the processes pertaining to this tier^{2–7}. Because their mean species duration is estimated at 2.3–2.6 Myr (refs 1, 2), primary (20–400 kyr) Milankovitch variations cannot be held responsible for mammal speciation and extinction^{4,9,10}. It has been suggested that amplitude changes of climatic oscillations could have played a part in explaining turnover^{3,10}, but until now no efforts have been made systematically to compare such amplitude variations with turnover in long and well-dated records.

We compiled a data set of more than 200 rodent assemblages from Central Spain (see Supplementary Notes and Supplementary Table 1) and analysed it in terms of turnover. The record is exceptionally long

(22 Myr), largely continuous, dense and well dated. We focused on rodents, because screen sieving allows the collection of large amounts of easily identifiable dental elements. The studied fossils originate from fluvio-lacustrine sections in the Madrid, Calatayud-Daroca and Teruel basins, and amount to over 80,000 isolated molars, which were identified at the species and lineage level (132 lineages, pseudo-extinctions ignored). All studied localities are positioned in stratigraphic sections, a large number of which have been tied to the geomagnetic polarity timescale by first-order correlations. The complete temporal sequence was calibrated to the new astronomically tuned timescale for the Neogene¹¹ (Supplementary Fig. 1). The use of this new timescale is of crucial importance because it allows a direct comparison with time series of the Earth's orbital parameters outlined above.

Randomization procedures were used to capture uncertainties in the ages of localities (by the generation of a series of equally probable age models) and of first and last rodent appearances, resulting in 1,000 equally probable time series. Sample size effects on the presence or absence of rodents were observed to be fairly small and were further reduced by inferring lineage presence on a range-through basis and by excluding the rarest occurrences (see Methods, Supplementary Notes and Supplementary Figs 4 and 5). We chose 0.1 Myr as our basic time unit and calculated a mean time series of standing diversity, and mean time series of total and per-taxon rates of origination, extinction and their aggregate (turnover; Figs 1 and 3b; Supplementary Figs 2 and 3). Re-entry by migration was assumed for lineages with an estimated absence of more than 1.0 Myr, resulting in 22 to 33 additional lineage segments, depending on the age model (see Methods and Supplementary Notes for details on age uncertainties and turnover calculations). It is not possible to make an accurate estimation of either the relative contributions of migration and speciation to the origination record or of local and true extinction to the extinction record, because of problems of poor dating and low resolution of time-equivalent records outside the study area. However, at the scale of Mammal Neogene (MN) European zones (mean duration of 1.3 Myr) we estimated that ~20% of both the entries and exits in the study region represent immigration or local extinction, respectively (excluding migration by re-entry). Almost 40% of the lineages are endemic to Spain. Their first- and last-occurrence ages and those of the remaining 40% can be considered to lie relatively close to the ages of speciation and true extinction.

The origination, extinction and turnover time series are characterized by distinct peaks spaced by ~1–2 Myr (Figs 1 and 3b;

¹Faculty of Geosciences, Department of Earth Sciences, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, The Netherlands. ²Faculty of Geology, Department of Paleontology, José Antonio Novais 2, Complutense University, Ciudad Universitaria, 28040 Madrid, Spain. ³National Museum of Natural History, Naturalis, Darwinweg 2, 2333 CR Leiden, The Netherlands. ⁴Faculty of Earth Sciences, University of Lyon 1, 2 rue Dubois, 69622 Villeurbanne Cedex, France. ⁵National Museum of Natural History, CSIC, Department of Paleobiology, c/o José Gutiérrez Abascal 2, 28006 Madrid, Spain. [†]Present address: Department of Earth and Environmental Sciences, Geophysics LMU Munich, Theresienstraße 41, D-80333 Munich, Germany.

Supplementary Figs 2 and 3). The rounded shape of the Early Miocene epoch turnover maxima can be attributed to the relatively large age uncertainties in this interval. Mostly, but not always, origination and extinction peaks coincide or have almost the same age. Using a bootstrap approach (see Methods and Supplementary Notes), we found that 11 entry, exit or turnover peaks were significant at the 99% level. At the 95%, 90% and 80% level, the total number of such significant events was 21, 29 and 33, respectively (Supplementary Table 2). We performed spectral analysis on the turnover rate series to test for million-year scale periodicities and found significant frequencies corresponding to periods of 2.4 or 2.5 Myr and 1.0 Myr (Fig. 2a, b). The individual extinction and origination spectra show a similar distribution of power (Supplementary Fig. 6). Most turnover maxima are associated with an increase in the proportion of short-lived species as indicated by the simultaneous occurrence of peaks in diversity and troughs in mean lifespan per 0.1 Myr (Fig. 1). The overall mean residence time is 2.0 Myr.

From astronomical calculations it is known that the amplitudes of the ~ 100 and 405 kyr eccentricity cycles occur superposed on long-period cycles with periods of 2.37 and 0.97 Myr (ref. 8; Figs 2a and 3e). In combination, these longer cycles produce prolonged intervals of low eccentricity that are either ~ 2.0 or ~ 2.8 Myr apart (Fig. 3e, blue stripes). The almost identical periods observed in the spectra of eccentricity and rodent turnover (Fig. 2a), as well as the recent documentation of the expression of the 2.37-Myr eccentricity cycle in continental and marine lithological sequences, and marine microfossil abundances^{12–14}, suggests that long-period eccentricity modulation had a significant influence on the rodent record. This notion is confirmed by cross-spectral analysis of the eccentricity envelope and per-taxon turnover series, which produces maximum and significant coherences at periods of 2.38 and 0.97 Myr, respectively (Fig. 2c). The phase spectrum, however, reveals that turnover is in anti-phase with the 2.37-Myr eccentricity cycle (0.05 ± 0.12 Myr, 2σ) and in phase

(0.06 ± 0.03 Myr) with the 0.97-Myr cycle (Figs 2c and 3c, e), excluding the possibility of a common forcing mechanism. On the other hand, eccentricity maxima occurring every 0.97-Myr that correlate to significant turnover events at 15.9–15.8, 14.8–14.7, 13.8–13.7, 9.0–8.7, 7.8–7.7, 4.3–4.1 and 3.3–3.1 Myr ago also coincide with distinct nodes (points of minimum variation) of the 1.2-Myr modulation cycle of obliquity⁸ (Fig. 3c, e, f; green stripes). The obliquity nodes are spaced 1.0 or 1.4 Myr apart in alternating fashion owing to a 2.4-Myr period (Figs 2b and 3f) produced by resonance with the equally long eccentricity cycle⁸. The appearance of the 0.98-Myr period in the turnover spectra can therefore also be explained by the 1.0-Myr spacing between obliquity nodes. Pairs of 1.0-Myr spaced combinations of eccentricity maxima and obliquity nodes occur once every 4 Myr on average at 20.6 and 19.6, 16.0 and 15.1, 13.7 and 12.7, 9.0 and 7.9, and 4.2 and 3.2 Myr ago (Fig. 3e, f). The cross spectrum of obliquity and turnover is consistent with obliquity influence, showing maximum and significant coherence at a frequency corresponding to a period of 1.17 Myr without any phase difference (0.01 ± 0.06 Myr; Fig. 2d). A faunal response to the 1.2-Myr obliquity cycle is to be expected because this cycle, in

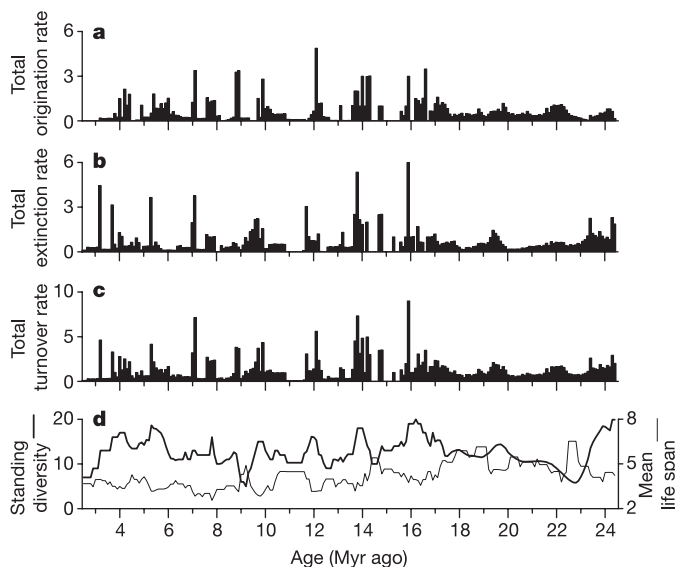


Figure 1 | Lineage turnover, diversity and mean lifespan per 0.1 Myr. The mean number of events and mean standing diversity are inferred from 1,000 equally probable series. The mean lifespan is based on an age model resulting in an average number of re-entering lineages, and calculated as the average lifespan of all lineages present. Origination ages before 24.4 Myr ago are based on the Oligocene rodent record from Central Spain and extinction ages after 2.5 Myr ago are based on the Pliocene–Pleistocene record of Southern Spain. **a**, Number of originations per 0.1 Myr. **b**, Number of extinctions per 0.1 Myr. **c**, Number of originations plus number of extinctions (turnover) per 0.1 Myr. **d**, Standing diversity (number of lineages) and mean lifespan (Myr).

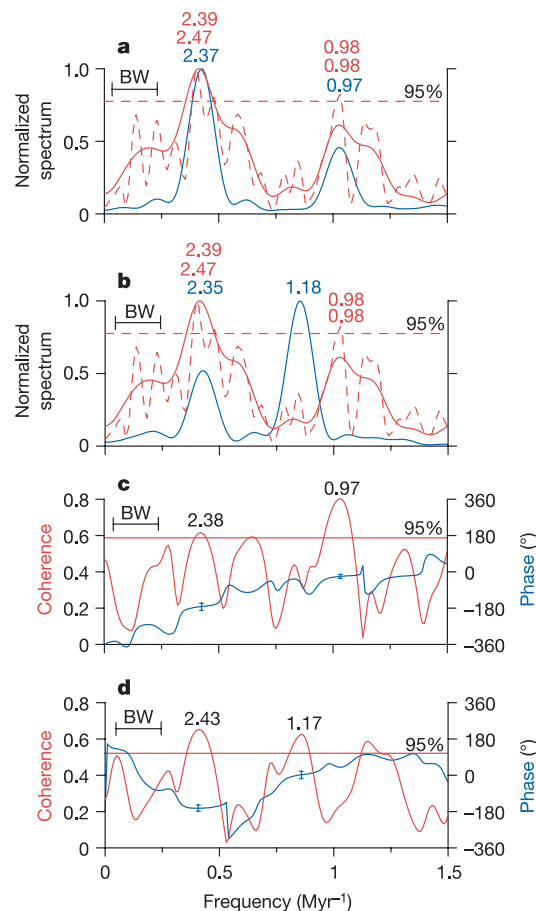


Figure 2 | Individual and cross spectra of turnover, eccentricity and obliquity. Methods used were Blackman–Tukey (solid lines), and CLEAN³⁰ (dashed lines). Numbers refer to significant periods (Myr). BW, bandwidth (for Blackman–Tukey). **a**, Red lines, per-taxon turnover. Blue line, eccentricity envelope (~ 100 kyr maxima). **b**, Red lines as in **a**. Blue line, obliquity envelope (41 kyr minima). **c**, Red line, coherence between per-taxon turnover and eccentricity maxima with 95% significance level for non-zero coherence. Blue line, phase difference; error bars (95% confidence level) are indicated for the relevant frequencies. **d**, Red line, coherence between per-taxon turnover and obliquity minima with 95% significance level for non-zero coherence. Blue line, phase difference; error bars (95% confidence level) are indicated for the relevant frequencies.

contrast to the 0.97-Myr eccentricity cycle, is well documented in marine sedimentary sequences^{12,14–18}.

How can reduced orbital variation at obliquity nodes and long-term eccentricity minima trigger biological turnover? The geological record suggests an important role for climatic change. Owing to the construction of high-resolution, astronomically tuned deep-sea records, many Cenozoic events of increasing $\delta^{18}\text{O}$ values (Oligocene isotope, Miocene isotope and related events) have been correlated to obliquity nodes with a mean spacing of 1.2 Myr^{14–18} (Fig. 3f, g; Supplementary Table 2). These million-year-scale events represent short (at most one to two hundred thousand years) episodes of

ice-sheet expansion and cooling superposed on the main Cenozoic-era cooling trend¹⁹ (Fig. 3h). The occurrence of these events is also related to eccentricity, because most of them seem to coincide with minima of the 405-kyr eccentricity cycle^{15–18}. The development of ice sheets can be explained by assuming a sustained reduction of ice melting during relatively cool high-latitude summers^{14,15}. It seems that almost all glacial events that are linked to the strongest combinations of an obliquity node and a 405-kyr eccentricity minimum (Fig. 3d) are associated with statistically significant turnover events in the well-dated part (after 17 Myr ago) of the rodent record (Supplementary Table 2).

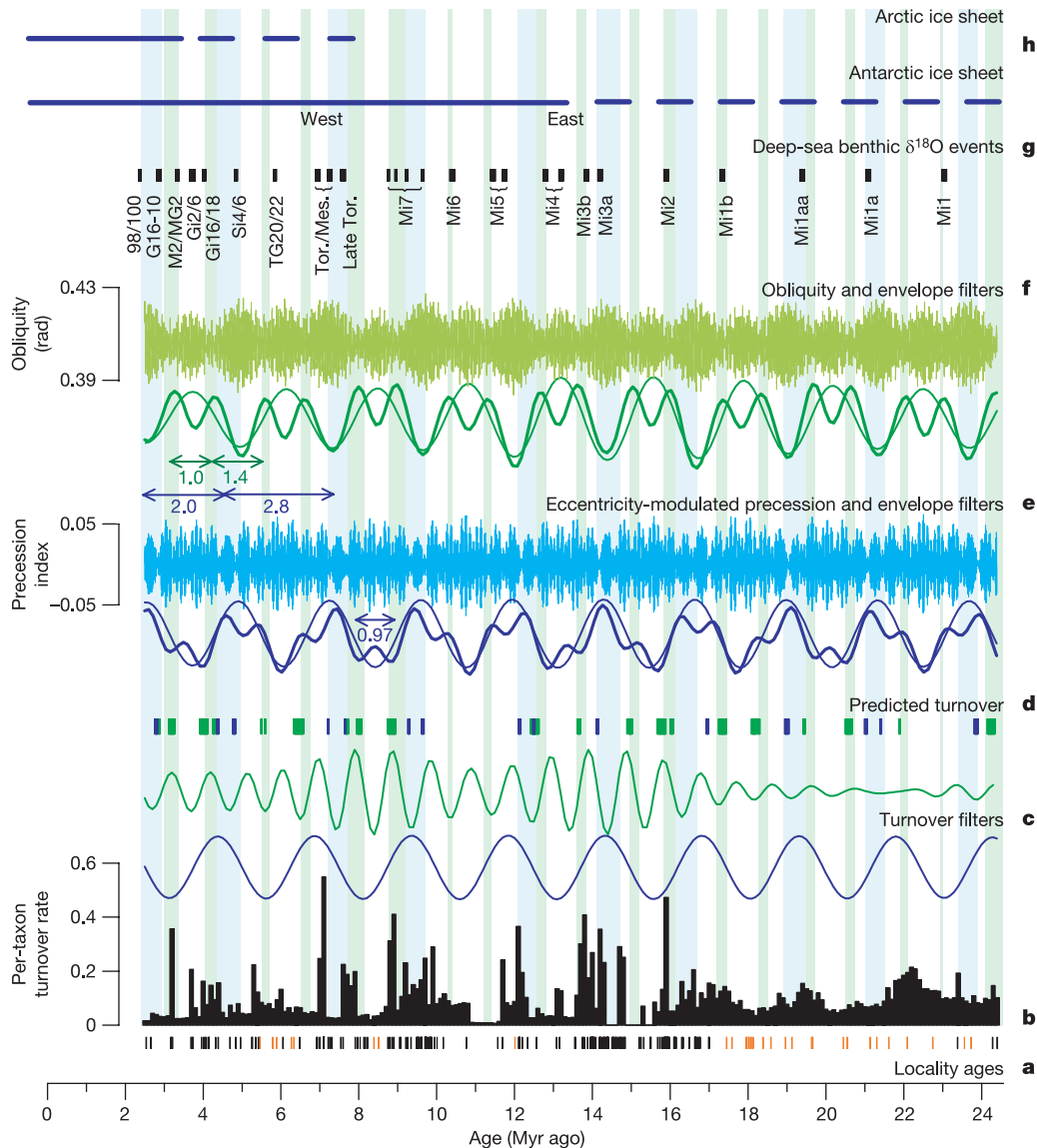


Figure 3 | Rodent turnover, astronomical parameters and climate. Blue stripes, eccentricity minima (mean spacing of 2.37 Myr), Green stripes, obliquity nodes (mean spacing of 1.2 Myr). **a**, Estimated ages of rodent localities. Black bars, from a single preferred correlation; orange bars, means after randomization (see Supplementary Notes and Supplementary Table 1). **b**, Per-taxon turnover rate (per 0.1 Myr). **c**, Filtered per-taxon turnover record. Blue, 99% significance filter, CLEAN method³⁰; green, filtered 0.97-Myr component, centred at $1.03 \pm 0.15 \text{ Myr}^{-1}$. **d**, Predicted times of turnover. Blue bars, mean eccentricity < 0.12 during preceding 100 kyr; green bars, mean 41-kyr obliquity minimum $> 0.396 \text{ rad}$ during preceding 400 kyr combined with mean eccentricity < 0.30 during preceding 100 kyr. **e**, Eccentricity-modulated precession and filters, CLEAN method³⁰. Upper

curve, eccentricity-modulated precession⁸; lower thin blue line, 99% significance filter of eccentricity maxima (reversed); lower thick blue line, 95% significance filter of eccentricity maxima (reversed). **f**, Obliquity and filters. Upper curve, obliquity⁸; lower thick green line, 90% significance filter of obliquity minima; lower thin green line, filtered 2.4-Myr component, centred at $0.43 \pm 0.1 \text{ Myr}^{-1}$. **g**, Major deep-sea $\delta^{18}\text{O}$ events. Mi, Miocene isotope; Tor., Tortonian; Mes., Messinian; TG, Thvera-Gauss; Si, Sidufjall; Gi, Gilbert; M, Mammoth; MG, Mammoth-Gauss; G, Gauss. **h**, Ice-sheet history¹⁹. Upper blue line, Arctic ice sheet; lower blue line, Antarctic ice sheet; continuous line, full-scale or permanent ice; dashed line, partial or ephemeral ice.

Positive feedback effects on the global climate of ice sheet expansion during obliquity nodes are expected to lead to increased continental cooling and aridification, which would then cause perturbations in terrestrial biota through reduced food availability. These deteriorating conditions, in turn, could lead to habitat fragmentation and extinction on the one hand, and to migration and lineage splitting on the other hand. The low proportions of humidity-indicating rodents and insectivores during Miocene obliquity nodes that are covered by rich and well-dated localities (16.1–15.8, 15.2–14.95, 13.85–13.6, 9.2–8.75, 8.15–7.7 Myr ago) confirm the presence of temporarily drier conditions (Supplementary Fig. 7). The lithological equivalents of these intervals are dominated by (or contain the transition to) red bed units, which represent the drier phases in the red bed–limestone alternations that are characteristic for the study area^{20–22}. In contrast, Pliocene obliquity nodes correlate with limestone-rich intervals²³, and are not associated with distinct positive or negative peaks in the proportion of humidity-adapted small mammals (5.3–2.5 Myr ago; Supplementary Fig. 7). Unfortunately, there are few high-resolution, well-dated Miocene–Pliocene humidity proxy records straddling obliquity nodes with which to make a comparison. Pliocene pollen assemblages from northwestern Africa associated with oxygen isotope stages 130 and 134 (Mammoth-Gauss 2 event; Fig. 3g) at 3.3 Myr ago²⁴ point to drier conditions for this region during the associated obliquity node. High carbonate levels in the Mediterranean sea²⁵ during the Miocene-isotope-5 $\delta^{18}\text{O}$ event at 11.6–11.4 Myr ago¹⁴ also seem to confirm the occurrence of drier regional conditions during Miocene nodes of obliquity.

The second dominant type of turnover occurs during eccentricity minima with a mean spacing of 2.37 Myr before 6 Myr ago, and also seems to be associated with climate-induced changes. Closer inspection of the four best-dated turnovers between 17 and 7 Myr ago (Fig. 3b) indicates that these tend to be concentrated at the fringes of these minima, where they correspond to strong 405-kyr eccentricity minima (Fig. 3d, e). A number of major $\delta^{18}\text{O}$ events are associated with this type of orbital configuration, for example, the Miocene-isotope-3a event at 14.2 Myr ago, peaks associated with the Miocene-isotope-7 event between 9.6–8.8 Myr ago, and two strong Tortonion-Messinian glaciations at 7.3–7.2 and 7.0–6.9 Myr ago (Fig. 3g; Supplementary Table 2). Also the events of Miocene-isotope-1a at 21.1 Myr ago, Miocene-isotope-1aa at 19.4 Myr ago, Sidufjall 4/6 at 4.85 Myr ago, and the glacial interval associated with isotope stages 98/100 at 2.4 Myr ago are associated with this configuration, indicating that low-amplitude eccentricity conditions favour global cooling and ice build-up even in the absence of low-amplitude obliquity changes. But in contrast to obliquity nodes, the proportions of humidity-adapted rodents and insectivores are high during eccentricity minima (Supplementary Figs 2, 3 and 7). The highest proportions occur at 16.7, 14.4–14.3, 12.1, 9.7–9.6 and 7.3 Myr ago (Supplementary Fig. 7), when 405-kyr minima are combined with high-amplitude obliquity. This observation is in accordance with our interpretation of more arid conditions during low-amplitude obliquity. Most of the wet-adapted small mammals that enter the record at these moments are immigrants with Central European source areas. Their southward expansions during periods of low eccentricity (mean spacing of 2.37 Myr) can be interpreted as a direct, regional response to changes in the Mediterranean precipitation regime^{26,27}.

The presence of major carbonate units indicative of lake expansion at stratigraphical positions corresponding to 2.37-Myr eccentricity minima^{20–22} is consistent with the higher humidity levels at these intervals as inferred from the small mammals. Moreover, detailed cyclostratigraphical analysis shows the expected weakening in the expression of predominantly precession-controlled sedimentary cycles during 405-kyr eccentricity minima that coincide or closely coincide with the 2.37-Myr minima^{13,22}. Recent climate modelling experiments show that western Mediterranean summers are warmer

and drier during a circular Earth orbit (zero eccentricity and no precessional effects) than during a precession maximum (at maximum eccentricity), but cooler and more humid than during a precession minimum (at maximum eccentricity; E. Tüenter, personal communication and ref. 28). These results suggest that the prolonged absence of extreme summer conditions—characterized by strong evaporation—that normally occur during strong precession minima favours lake extension and leads to an increase of both wet-adapted small-mammal lineages and total diversity during minimum eccentricity (Fig. 1d). It can be argued that regional biogeographical and genetic processes raise turnover rates even further, because generalism and vagility (the ability to disperse), properties that would normally contribute to spatial and genetic re-mixing during successive climate swings, are temporarily less selected for during prolonged intervals without extreme climates. At the same time, specialization trends are expected to continue for a longer time during such intervals and to have a higher probability of leading to successful speciation^{10,29}. These arguments are consistent with the observation that Pleistocene turnover rates are relatively low despite the occurrence of frequent and strong climatic events^{9,10}.

The apparent correlation between orbital configurations, climatic events and rodent turnover strongly suggests that the latter is controlled by astronomically forced climate change. Most intervals of no correlation either correspond to Miocene intervals during which amplitudes and durations of the orbital parameters do not reach specific threshold values (Fig. 3d; Supplementary Table 2), or to the Pliocene, for which additional relationships between long-term astronomical cycles and climate have been proposed, including a forcing of bipolar ice-sheet growth by maximum obliquity variation¹² (Fig. 3f, g).

The postulated astronomical hypothesis for species turnover provides a crucial missing piece in the puzzle of mammal species- and genus-level evolution, and provides a probable mechanism to explain third-tier processes. The new hypothesis is consistent with the Turnover Pulse hypothesis³ in the sense that lineage events tend to cluster and that the clusters correlate to abiotic events. In addition, the astronomical hypothesis for turnover offers a plausible explanation for the characteristic duration of ~2.5 Myr of the mean species lifespan in mammals, and may explain similar durations in other biological groups as well.

METHODS

Age uncertainties surrounding poorly constrained localities and first- and last-appearance datums were addressed by generating 1,000 equally probable, equally spaced (0.1 Myr) time series ('age models') for origination, extinction and turnover (sum of origination and extinction) for the entire study interval. The final target series used consisted of the means of all 1,000 series. The 1,000 time series were produced by combining 50 locality age models with 20 sets of random draws from the uncertainty intervals preceding first appearances or succeeding last appearances. These intervals were calculated on the basis of sample size by calculating the probability of finding a taxon in preceding (or succeeding) localities given the proportion in the locality where it was first (or last) recorded, and using a cut-off value of 80% (ref. 7). The same procedures were applied to range gaps. These were defined as true gaps if the midpoints of the uncertainty intervals differed by 1.0 Myr or more, in which case the lineage was split for the subsequent analyses. In the cases of gaps less than 1.0 Myr, the taxon was considered to have remained present (range-through approach; see Supplementary Notes for more details).

Clustering of events was tested by running a separate bootstrap analysis for each of the 1,000 sets of origination, extinction and turnover. The analyses involved the comparison of observed numbers of events per moving 0.3-Myr interval (the chosen bootstrap test statistic) with the corresponding numbers of events from 1,000 rounds of randomly reshuffling 0.1-Myr level first- and last-appearance ages of all lineages across the presence intervals of their families. The final significance levels were calculated as the average levels over the 1,000 time series. Each individual level was calculated as the midpoint of that part of the cumulative relative frequency distribution (of the number of events resulting from the 1,000 rounds of reshuffling) that corresponded to the observed number of events. Significant intervals are included in Supplementary Table 2.

Received 16 January; accepted 11 August 2006.

1. Alroy, J. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology* **26**, 707–733 (2000).
2. Vrba, E. S. & DeGusta, D. Do species populations really start small? New perspectives from the Late Neogene fossil record of African mammals. *Phil. Trans. R. Soc. Lond. B* **359**, 285–293 (2004).
3. Vrba, E. S. in *Paleoclimate and Evolution, with Emphasis on Human Origins* (eds Vrba, E. S., Denton, G. H., Partridge, T. C. & Burckle, L. H.) 385–424 (Yale, New Haven, 1995).
4. Barnosky, A. D. Distinguishing the effects of the Red Queen and Court Jester on Miocene mammal evolution in the Northern Rocky Mountains. *J. Vertebr. Paleontol.* **21**, 172–185 (2001).
5. Prothero, D. R. & Heaton, T. H. Faunal stability during the Early Oligocene climatic crash. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **127**, 257–283 (1996).
6. Alroy, J., Koch, P. L. & Zachos, J. C. in *Deep Time: Paleobiology's Perspective* (eds Erwin, D. H. & Wing, S. L.) 259–288 (Paleontological Society, Lawrence, 2000).
7. Barry, J. *et al.* Faunal and environmental change in the Late Miocene Siwaliks of Northern Pakistan. *Paleobiol. Memoirs* **3**, suppl. to *Paleobiology* **28**, 1–71 (2002).
8. Laskar, J. *et al.* A long term numerical solution for the insolation quantities of the Earth. *Astron. Astrophys.* **428**, 261–285 (2004).
9. Bennett, K. D. *Evolution and Ecology: the Pace of Life* (Cambridge Univ. Press, 1997).
10. Dynesius, M. & Jansson, R. Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. *Proc. Natl Acad. Sci. USA* **97**, 9115–9120 (2000).
11. Lourens, L. J., Hilgen, F. J., Laskar, J., Shackleton, N. J. & Wilson, D. in *A Geological Time Scale 2004* (eds Gradstein, F. M., Ogg, J. G. & Smith, A. G.) 409–440 (Cambridge Univ. Press, 2004).
12. Lourens, L. J. & Hilgen, F. J. Long-periodic variations in the Earth's obliquity and their relation to third-order eustatic cycles and late Neogene climate change. *Quat. Int.* **40**, 43–52 (1997).
13. Abdul Aziz, H., Krijgsman, W., Hilgen, F. J., Wilson, D. S. & Calvo, J. P. An astronomical polarity timescale for the late middle Miocene based on cyclic continental sequences. *J. Geophys. Res.* **108** (B3), 2159–2175 (2003).
14. Turco, E., Hilgen, F. J., Lourens, L. J., Shackleton, N. J. & Zachariasse, W. J. Punctuated evolution of global climate cooling during the late Middle to early Late Miocene: high-resolution planktonic foraminiferal and oxygen isotope records from the Mediterranean. *Paleoceanography* **16**, 405–423 (2001).
15. Zachos, J., Shackleton, N. J., Revenaugh, J. S., Pälike, H. & Flower, B. P. Climate response to orbital forcing across the Oligocene–Miocene boundary. *Science* **292**, 274–278 (2001).
16. Wade, B. S. & Pälike, H. Oligocene climate dynamics. *Paleoceanography* **19**, doi:10.1029/2004PA001042 (2004).
17. Westerhold, T., Bickert, U. & Röhl, U. Middle to late Miocene oxygen isotope stratigraphy of ODP site 1085 (SE Atlantic): new constraints on Miocene climate variability and sea-level fluctuations. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **217**, 205–222 (2005).
18. Abels, H. A. *et al.* Long-period orbital control on middle Miocene global cooling: integrated stratigraphy and astronomical tuning of the Blue Clay Formation on Malta. *Paleoceanography* **20**, doi:10.1029/2004PA001129 (2005).
19. Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
20. Daams, R. *et al.* Stratigraphy and sedimentology of the Aragonian (early to middle Miocene) in its type area (north-central Spain). *Newslett. Stratigr.* **37**, 103–139 (1999).
21. Alonso Zarza, A. M. & Calvo, J. P. Palustrine sedimentation in an episodically subsiding basin: the Miocene of the northern Teruel Graben. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **160**, 1–21 (2000).
22. Abdul Aziz, H., van Dam, J. A., Hilgen, F. J. & Krijgsman, W. Astronomical forcing in Upper Miocene continental sequences: implications for the Geomagnetic Polarity Time Scale. *Earth Planet. Sci. Lett.* **222**, 243–258 (2004).
23. Mein, P., Moissenet, E. & Adrover, R. Biostratigraphie du Néogène supérieur de Teruel. *Paleont. Evol.* **23**, 121–139 (1990).
24. Leroy, S. & Dupont, L. Development of vegetation and continental aridity in northwestern Africa during the Late Pliocene: the pollen record of ODP Site 658. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **109**, 295–316 (1994).
25. Hilgen, F. J., Abdul Aziz, H., Krijgsman, W., Raffi, I. & Turco, E. Integrated stratigraphy and astronomical tuning of the Serravallian and lower Tortonian at Monte dei Corvi (Middle-Upper Miocene, northern Italy). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **199**, 229–264 (2003).
26. van Dam, J. A. & Weltje, G. J. Reconstruction of the late Miocene climate of Spain using rodent paleocommunity successions: an application of end-member modelling. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **151**, 267–305 (1999).
27. van Dam, J. A. Anourosoricini (Soricidae, Mammalia) from the Mediterranean Region: a pre-Quaternary example of recurrent climate-controlled North-South range shifting. *J. Paleontol.* **78**, 741–764 (2004).
28. Tüenter, E. *Modeling Orbital Induced Variations in Circum-Mediterranean Climate*. PhD thesis, Utrecht Univ. (2004).
29. Sheldon, P. R. Plus ça change—a model for stasis and evolution in different environments. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **127**, 209–227 (1996).
30. Heslop, D. & Dekkers, M. J. Spectral analysis of unevenly spaced climatic time series using CLEAN; signal recovery and derivation of significance levels using a Monte Carlo simulation. *Phys. Earth Planet. Inter.* **130**, 103–116 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Netherlands Organisation for Scientific Research and the Comunidad Autónoma de Madrid for financial support, and the Ministerio de Educación y Ciencia (Madrid), the Diputación General de Aragón (Zaragoza), and the Fundación Conjunto Paleontológico (Dinópolis, Teruel) for their support with regard to fieldwork campaigns. We also thank H. Abels, H. Dijkstra, D. Heslop, B. Engesser, R. Hengeveld, K. Kuiper, D. Mayhew, E. Sanz, E. Tüenter and G. J. Weltje for discussions and help. We thank I. Lacombe for supplying unpublished data and we are grateful to the late R. Daams for his important contribution to the Central Spanish rodent database.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.A.v.D. (jdam@geo.uu.nl).

LETTERS

Low-coverage vaccination strategies for the conservation of endangered species

D. T. Haydon¹, D. A. Randall^{2,3}, L. Matthews⁴, D. L. Knobel^{3,4}, L. A. Tallents^{2,3}, M. B. Gravenor⁵, S. D. Williams^{2,3}, J. P. Pollinger⁶, S. Cleaveland⁴, M. E. J. Woolhouse⁷, C. Sillero-Zubiri^{2,3}, J. Marino^{2,3}, D. W. Macdonald² & M. K. Laurenson^{3,4,8}

The conventional objective of vaccination programmes is to eliminate infection by reducing the reproduction number of an infectious agent to less than one¹, which generally requires vaccination of the majority of individuals. In populations of endangered wildlife, the intervention required to deliver such coverage can be undesirable and impractical²; however, endangered populations are increasingly threatened by outbreaks of infectious disease for which effective vaccines exist^{3,4}. As an alternative, wildlife epidemiologists could adopt a vaccination strategy that protects a population from the consequences of only the largest outbreaks of disease. Here we provide a successful example of this strategy in the Ethiopian wolf, the world's rarest canid⁵, which persists in small subpopulations threatened by repeated outbreaks of rabies introduced by domestic dogs⁶. On the basis of data from past outbreaks, we propose an approach that controls the spread of disease through habitat corridors between subpopulations and that requires only low vaccination coverage. This approach reduces the extent of rabies outbreaks and should significantly enhance the long-term persistence of the population. Our study shows that vaccination used to enhance metapopulation persistence through elimination of the largest outbreaks of disease requires lower coverage than the conventional objective of reducing the reproduction number of an infectious agent to less than one¹.

The use of safe and effective vaccination can have a vital role in managing infectious disease in wildlife populations. For logistic, economic and ethical reasons, however, it will always be desirable to minimize the number of animals to be vaccinated. We distinguish between two different uses of vaccination: one focused on eliminating disease from a population, and another focused on protecting an endangered population from extinction. The design of vaccination programmes to eliminate infectious disease from populations has received much attention^{1,7,8}, and usually requires vaccinating a proportion of the population upward of $1-1/R_0$, where R_0 is the reproduction number of the infectious agent¹. A conceptually distinct approach is to assume that wildlife populations can tolerate limited outbreaks of disease, but their viability is threatened by large outbreaks that could reduce their size to below a minimally viable threshold^{9,10}. Targeted vaccination could be then used to curtail the largest and most damaging outbreaks, while reducing the proportion of individuals required to be vaccinated. Here we examine the effectiveness of such a strategy in conserving populations of Ethiopian wolves (*Canis simensis*) threatened by outbreaks of rabies, a fatal viral disease of mammals.

Ethiopian wolves in the Bale Mountains persist in several subpopulations connected by narrow corridors of habitat¹¹ (Fig. 1). Within these subpopulations, two large outbreaks of rabies were detected in 1992 and 2003 (ref. 12), and rabies was the suspected cause of a population crash in 1991 (refs 12–14); canine distemper was also suspected to have infected wolves in 1993 (ref. 15). These repeated introductions of infection into the wolf population from a domestic dog reservoir⁶, together with permission to mount a reactive vaccination campaign in response to one such outbreak, have provided a rare opportunity to test experimentally a vaccination control strategy in an endangered species. Here we briefly review the

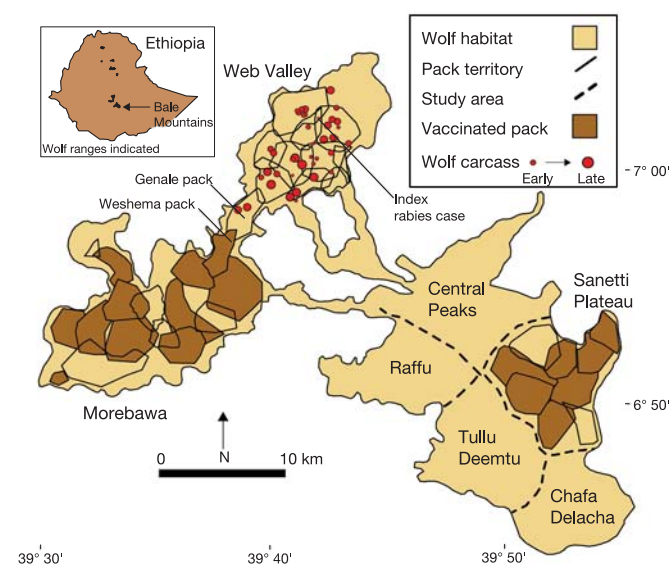


Figure 1 | Known distribution of Ethiopian wolf packs in three subpopulations in the Bale Mountains. Shown is the configuration of territories in the Web Valley, Sanetti Plateau and Morebawa, including the Genale and Weshema packs located in and immediately beyond the corridor linking the Web Valley with Morebawa. Packs are present in Central Peaks, Raffu, Chafa Delacha and Tullu Deemtu areas, but territory boundaries are not known with any precision. Filled red circles indicate the location of carcasses found in the 2003 outbreak; circle size indicates the relative timing of carcass recovery. Filled polygons indicate vaccinated packs. Inset shows the current distribution of the species throughout Ethiopia.

¹Division of Environmental and Evolutionary Biology, University of Glasgow, Glasgow G12 8QQ, UK. ²Wildlife Conservation Research Unit, University of Oxford, Tubney House, Oxford OX13 5QL, UK. ³Ethiopian Wolf Conservation Programme, PO Box 215, Robe, Bale, Ethiopia. ⁴Wildlife and Emerging Diseases Section, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian EH25 9RG, UK. ⁵Institute of Life Science, School of Medicine, Swansea University, Singleton Park, Swansea, SA2 8PP, UK. ⁶Conservation Genetics Resource Center, University of California, 621 Charles E. Young Drive South, Los Angeles, California 90095, USA. ⁷Centre for Infectious Diseases, University of Edinburgh, Ashworth Laboratories, Kings Buildings, West Mains Road, Edinburgh EH9 3JF, UK. ⁸Frankfurt Zoological Society, PO Box 14935, Arusha, Tanzania.

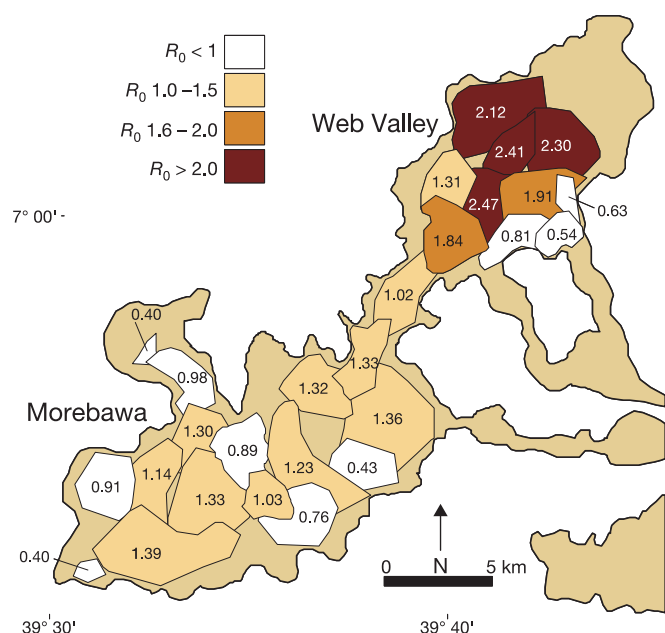


Figure 2 | R_0 map of the Web Valley and Morebawa subpopulations. Mixing is assumed to be intermediate (between pack transmission is 10% of within pack transmission). R_0 was predicted from the application of per capita transmission rates estimated from the fit of the SEIR model to data from the Web Valley, calculated by direct simulation assuming a single index case arose within each pack.

outbreak and the adopted vaccination strategy, and then describe and parameterize an epidemiological model that predicts the potential course of the outbreak without vaccination. Finally, we extend the model to explore the increase in population persistence times resulting from vaccinating in subpopulations as compared with vaccinating only in packs occupying corridor habitats between subpopulations. Additional data and analyses are given in the Supplementary Information.

In August 2003, the Bale wolf population comprised 200–250 individuals (aged >1 yr) in 36 packs associated with known territory

locations. In the Web Valley, roughly 95 wolves resided in ten packs. Between August 2003 and January 2004, an outbreak of rabies resulted in the deaths of 72 (76%) of these wolves¹². Disease spread outward from the putative index case, and a total of 38 carcasses were recovered from all territories in the Web Valley, suggesting a carcass detection probability of ~50% (38/72; Fig. 1). A parenteral reactive vaccination programme implemented in November was designed to prevent transmission between subpopulations, while limiting the number of wolves handled. The campaign initially targeted packs immediately beyond the corridor connecting the Web Valley and Morebawa subpopulations (Genale pack, Fig. 1) and then gradually moved to packs further away from the disease front (Fig. 1). Final coverage averaged over the whole Morebawa subpopulation was 37.5%. Subsequent monitoring suggested a maximum of seven deaths out of 105 wolves in the Morebawa subpopulation, all from a single front-line territory (Weshema pack; Fig. 1). Although carcasses continued to be recovered in the Web Valley, the epidemic made no further incursion into the vaccination zone.

To determine whether the vaccination programme itself limited the severity of the outbreak we developed a spatially explicit demographically stochastic susceptible–exposed–infectious–removed (SEIR) model⁹. The model was parameterized, taking into account pack and population age structure and the final epidemic size in the Web Valley (Supplementary Table S2 and Fig. S1). The maximum likelihood estimate of R_0 depended on the assumed ratio of between (β_b) to within (β_w) pack transmission (Supplementary Table S3). The results suggested, however, that the observed pattern of mortality across packs was most consistent with an intermediate to high ratio of β_b/β_w (see Supplementary Information), and here we present results assuming that $\beta_b = 0.1\beta_w$ (further results and genetic data in support of this choice are reported in the Supplementary Information). The maximum likelihood estimate of R_0 for the 2003 outbreak was 2.4 (95% confidence intervals: 1.7–3.4; Supplementary Table S3). A risk map constructed from predicted values for R_0 for epidemics starting in different packs reflects the importance of pack composition and configuration (Fig. 2). In Morebawa, pack size averaged 6.1 individuals, as compared with 8.8 in the Web Valley, and predicted R_0 values were correspondingly lower.

By repeatedly simulating outbreaks in subpopulations structured as in 2003, the model suggested that there was a 40% chance that

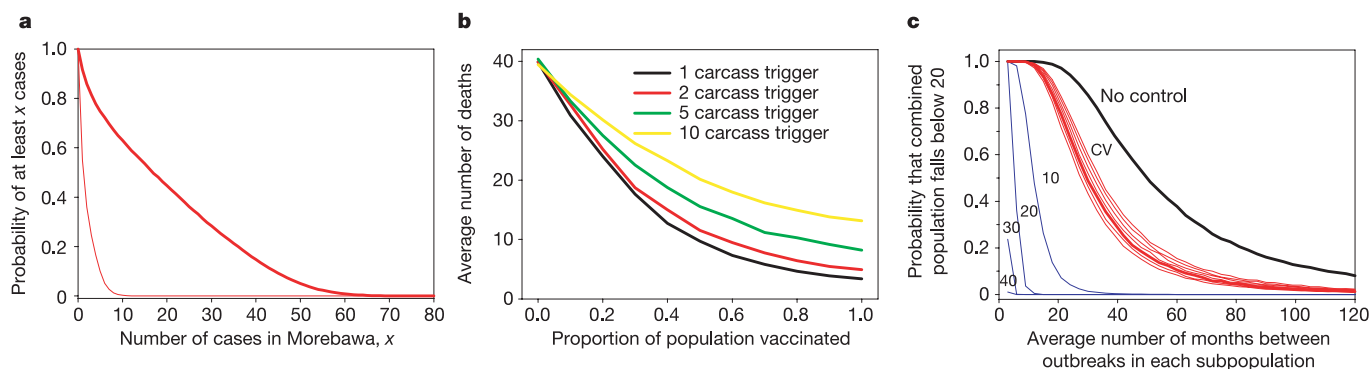


Figure 3 | Model projections. **a**, Probability that a rabies outbreak starting with a single index case in the central Web Valley goes on to cause a subsequent outbreak in the subpopulation in Morebawa, assuming the subpopulation to be fully susceptible and unvaccinated. Results are from an intermediately mixed model ($R_0 = 2.4$) in which per capita transmission rates between packs were 10% of within packs (thick red line). Thin red line indicates the outcome of an equivalent simulation that includes the effect of the vaccination programme as implemented during the outbreak. **b**, Effects of reactive vaccination in the Web Valley subpopulation on total epidemic size implemented after the deaths of different numbers of wolves. Vaccination was assumed to be instantaneously protective when administered to unexposed individuals (see Supplementary Information),

but not effective when administered to incubating individuals. **c**, Probability of catastrophic metapopulation reduction occurring over a 20-yr period as a function of increasing rates of disease introduction into each subpopulation. The model assumes three subpopulations, linked by habitat and migration, that support a maximum of 100 individuals each. Black line, no control; red lines, corridor vaccination (CV) that reduces the probability of an epidemic in one subpopulation spreading to another from 0.25 to a range of possible values rising in increments of 0.02 from 0.04 to 0.16; blue lines, reactive vaccination implemented after the death of ten individuals with coverages of 10, 20, 30 and 40%, as indicated (see Supplementary information for further details of the model).

rabies epidemics arising from a single index case would fade out with less than ten (and usually less than four) individuals becoming infected. Once they exceeded this number, however, the epidemic would almost certainly go on to be large. Results from the model indicated that the probability of infection passing through the corridor into Morebawa increased from 0.08 with vaccination (as implemented during 2003) to 0.25 in the absence of vaccination (Supplementary Table S3). The upper ninety-fifth percentile interval estimates for final outbreak size in the Morebawa subpopulation started through this route increased from 8 wolves with vaccination to 41 without (Fig. 3a). Given how close the estimated R_0 values are to the unit threshold for packs in this subpopulation (Fig. 2), small increases in wolf density could result in sharp rises in the probability of even more damaging outbreaks.

Assuming a randomly located index case in the Web Valley, the model predicted an average time delay between the first death and an individual becoming infected in the Genale pack in the corridor to Morebawa at 74 d. The corresponding values for the second, fifth and tenth deaths were 66 d (lower eightieth percentile interval: 33 d), 51 d (21 d) and 38 d (10 d), respectively. This suggests that even if carcass detection rates fall as low as 20%, there would still be sufficient time in which to implement a reactive corridor vaccination campaign triggered by the detection of two carcasses.

By the time permission to vaccinate had been granted, the outbreak was considered too far advanced to protect the Web Valley subpopulation. Modification of the model to include reactive vaccination in an affected subpopulation triggered after the death of different numbers of individuals suggested, however, that vaccination would remain beneficial even after 10% of the subpopulation had died from rabies (Fig. 3b).

These analyses show that the extent of rabies outbreaks can be limited by low-coverage reactive vaccination strategies, but does such action mitigate the risk of extinction in the longer term? We used a population viability analysis¹⁶ (PVA) to demonstrate the considerable risk posed by rabies outbreaks to these subpopulations and the substantial benefits of low-coverage reactive vaccination for the persistence of the metapopulation as a whole. The PVA assumed that habitat corridors between subpopulations act as conduits for disease transmission (see Supplementary Information) but also facilitate migration and recovery after epidemics¹⁷. The probability of catastrophic reduction of the metapopulation (defined here as the combined metapopulation falling below 20 individuals) at any time over a 20-yr period fell in response to implementation of reactive corridor vaccination, and even more quickly as reactive core vaccination strategies were adopted that targeted 40, 30, 20 and as little as 10% of the affected subpopulation (Fig. 3c). For example, the PVA predicted that, if rabies virus was introduced at an average rate of once every 5 yr into each subpopulation, corridor vaccination would reduce the probability of catastrophic metapopulation reduction almost fourfold from 0.38 to 0.10, and that reactive core vaccination of only 10% of individuals would reduce this probability to <0.001 .

Aside from specific recommendations for this population (see Supplementary Information), our analyses underline the general importance of baseline ecological data, surveillance and detailed quantitative contingency planning in the management of epidemics^{18–22}. High-quality demographic data enable interventions to be targeted, effective monitoring is essential for the early detection of suspected disease outbreaks, and appropriately calibrated trigger points minimize unnecessary interventions and facilitate the timing of decisive and effective action once an outbreak occurs. Taken together, these steps enable the threat of infectious disease to be managed through a programme of minimally invasive but demographically significant interventions.

Although preventative vaccination of reservoir hosts can reduce the frequency of stochastic spill-over infections into wildlife⁴, the risk of outbreaks in unvaccinated wildlife populations cannot be eliminated, particularly when limited resources restrict the extent

and coverage of reservoir vaccination programmes. Vaccination and handling of African canids has generated considerable controversy in the past^{2,23}, but our analysis provides strong evidence that targeted low-coverage and less-invasive reactive vaccination strategies can be effective in curtailing disease outbreaks and enhance the long-term persistence of endangered populations. However, progress is required in the development of protocols for more logistically feasible and cost-effective vaccine delivery methods such as oral vaccination^{24–27}, and policy-makers and conservation practitioners must be provided with epidemiologically sound, practical advice with which to develop contingency plans. Greater knowledge of the spatial ecology and social organization of other endangered species is likely to be fundamental to the development of practical and effective solutions to enduring threats to their persistence posed by infectious disease.

METHODS

Vaccination. Details of wolf monitoring and vaccination procedures are detailed in the Supplementary Information.

Models. We used a conventional SEIR model, which assumes demographically stochastic dynamics. The average infectious period was assumed to be 5 d (ref. 28), and an average incubation period of 12 d was fitted to match the duration of the outbreak. The model was fitted to data from the Web Valley, and then used to predict the probability of spread between subpopulations and the impact of such outbreaks in these subpopulations. Further details of the epidemiological model and its parameterization are supplied in the Supplementary Information.

The PVA model used demographic parameters from long-term monitoring studies, and probabilities of between subpopulation spread of infection and simulated distributions of outbreak sizes predicted by the model. Further details of this model are supplied in the Supplementary Information, together with a description of the genetic methods and analyses that indicated only limited movement of infected individuals.

Received 14 June; accepted 15 August 2006.

- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans, Dynamics and Control* (Oxford Univ. Press, Oxford, 1992).
- Woodroffe, R. Assessing the risks of intervention: immobilization, radio-collaring and vaccination of African wild dogs. *Oryx* **35**, 234–244 (2001).
- Woodroffe, R., Cleaveland, S., Courtenay, O., Laurenson, K. & Artois, M. in *Biology and Conservation of Wild Canids* (eds Macdonald, D. W. & Sillero-Zubiri, C.) 123–142 (Oxford University Press, Oxford, 2004).
- Daszak, P., Cunningham, A. A. & Hyatt, A. D. Emerging infectious diseases of wildlife: threats to biodiversity and human health. *Science* **287**, 443–449 (2000).
- Marino, J. Threatened Ethiopian wolves persist in small isolated Afroalpine enclaves. *Oryx* **37**, 62–71 (2003).
- Randall, D. A. *et al.* An integrated disease management strategy for the control of rabies in Ethiopian wolves. *Biol. Conserv.* **131**, 151–162 (2006).
- Hethcote, H. W. An immunization model for a heterogeneous population. *Theor. Pop. Biol.* **14**, 338–349 (1978).
- Ball, F. G. & Lyne, O. D. Optimal vaccination policies for stochastic epidemics among a population of households. *Math. Biosci.* **177–78**, 333–354 (2002).
- Haydon, D. T., Laurenson, M. K. & Sillero-Zubiri, C. Integrating epidemiology into population viability analysis: managing the risk posed by rabies and canine distemper to the Ethiopian wolf. *Conserv. Biol.* **16**, 1372–1385 (2002).
- Vial, F., Cleaveland, S., Rasmussen, G. & Haydon, D. T. Development of vaccination strategies for the management of rabies in African wild dogs. *Biol. Conserv.* **131**, 180–192 (2006).
- Sillero-Zubiri, C. & Marino, J. in *Canids: Foxes, Wolves, Jackals and Dogs. Status Survey and Conservation Action Plan* (eds Sillero-Zubiri, C., Hoffmann, M. & Macdonald, D. W.) 167–173 (IUCN/SSC Canid Specialist Group, Gland, Switzerland, and Cambridge, UK, 2004).
- Randall, D. A. *et al.* Rabies in endangered Ethiopian wolves. *Emerg. Inf. Dis.* **10**, 2214–2217 (2004).
- Sillero-Zubiri, C., King, A. A. & Macdonald, D. W. Rabies and mortality in Ethiopian wolves (*Canis simensis*). *J. Wildl. Dis.* **32**, 80–86 (1996).
- Marino, J., Sillero-Zubiri, C. & Macdonald, D. W. Trends, dynamics and resilience of an Ethiopian wolf population. *Anim. Conserv.* **9**, 49–58 (2006).
- Laurenson, K. *et al.* Disease as a threat to endangered species: Ethiopian wolves, domestic dogs and canine pathogens. *Anim. Conserv.* **1**, 273–280 (1998).
- Morris, W. F. & Doak, D. F. *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis* (Sinauer, Sunderland, MA, 2002).
- Hess, G. Disease in metapopulation models: implications for conservation. *Ecology* **77**, 1617–1632 (1996).
- Ferguson, N. M. *et al.* Planning for smallpox outbreaks. *Nature* **425**, 681–685 (2003).

19. Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G. & Grenfell, B. T. Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136–142 (2003).
20. Keeling, M. J. *et al.* Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813–817 (2001).
21. Anderson, R. M. *et al.* Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. *Phil. Trans. R. Soc. Lond. B* **359**, 1091–1105 (2004).
22. Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science* **292**, 1155–1160 (2001).
23. Morell, V. Wildlife biology—dogfight erupts over animal studies in the Serengeti. *Science* **270**, 1302–1303 (1995).
24. Farry, S. C., Henke, S. E., Beasom, S. L. & Fearnleyhough, M. G. Efficacy of bait distributional strategies to deliver canine rabies vaccines to coyotes in southern Texas. *J. Wildl. Dis.* **34**, 23–32 (1998).
25. Knobel, D. L., Du Toit, J. T. & Bingham, J. Development of a bait and baiting system for delivery of oral rabies vaccine to free-ranging African wild dogs (*Lycaon pictus*). *J. Wildl. Dis.* **38**, 352–362 (2002).
26. Knobel, D. L., Liebenberg, A. & Du Toit, J. T. Seroconversion in captive African wild dogs (*Lycaon pictus*) following administration of a chicken head bait/SAG-2 oral rabies vaccine combination. *Onderstepoort J. Vet. Res.* **70**, 73–77 (2003).
27. Steelman, H. G., Henke, S. E. & Moore, G. M. Gray fox response to baits and attractants for oral rabies vaccination. *J. Wildl. Dis.* **34**, 764–770 (1998).
28. Foggin, C. M. *Rabies and Rabies Related Viruses in Zimbabwe: Historical, Virological, and Ecological*. PhD thesis, Univ. Zimbabwe (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Ethiopia's Wildlife Conservation Department, the Oromiya Rural Land and Natural Resources Administration Authority, and the

Bale Mountains National Park for permission to undertake this work. We thank the staff of the EWCP for field work; the WCD veterinarians F. Shiferaw and K. Argaw; C. Rupprecht and staff at CDC; T. Fooks and staff at the VLA; R. K. Wayne for the work in his genetics laboratory, supported in part by his grant from the NSF; and the IUCN/SSC Canid and Veterinary Specialist Groups for advice. D.T.H. acknowledges the award of a MacLagan Travel Grant from the Royal Society of Edinburgh in support of this research. Funding was provided by the Born Free Foundation, Frankfurt Zoological Society, the Wellcome Trust, Wildlife Conservation Network, Morris Animal Foundation, Conservation International and Siren UK.

Author Contributions D.T.H. undertook much of the model formulation and analysis, and coordinated the synthesis of ideas and information. D.A.R. generated and compiled much of the empirical and all of the genetic data. L.M. assisted with formulation and interpretation of the analyses. D.L.K. managed the implementation of the vaccination programme. L.A.T. provided data and analysis for spatial distribution of the wolf population. M.B.G. assisted with early formulation of the modelling work and interpretation of later model output. S.D.W. coordinated fieldwork for much of the period over which this study applies. J.P.P. was instrumental in overseeing the genetic analyses. S.C. assisted in formulation of the vaccination strategy and in writing the manuscript. M.E.J.W. suggested analyses and assisted in writing the manuscript. C.S.-Z. was responsible for funding and coordinating Ethiopian wolf research and conservation work, and provided data on the 1992 outbreak. J.M. provided essential data on wolf demography for use in the PVA model. D.W.M. conceived and supervised several doctoral studies that have underpinned the project since its inception. M.K.L. was responsible for designing and implementing the disease control strategy.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.T.H. (D.Haydon@bio.gla.ac.uk).

LETTERS

Transforming the architecture of compound eyes

Andrew C. Zelhof¹, Robert W. Hardy¹, Ann Becker¹ & Charles S. Zuker¹

Eyes differ markedly in the animal kingdom, and are an extreme example of the evolution of multiple anatomical solutions to light detection and image formation. A salient feature of all photoreceptor cells is the presence of a specialized compartment (disc outer segments in vertebrates, and microvillar rhabdomeres in insects), whose primary role is to accommodate the millions of light receptor molecules required for efficient photon collection. In insects, compound eyes can have very different inner architectures^{1–3}. Fruitflies and houseflies have an open rhabdom system, in which the seven rhabdomeres of each ommatidium are separated from each other and function as independent light guides. In contrast, bees and various mosquitoes and beetle species have a closed system, in which rhabdomeres within each ommatidium are fused to each other, thus sharing the same visual axis. To understand the transition between open and closed rhabdom systems, we isolated and characterized the role of *Drosophila* genes involved in rhabdomere assembly. Here we show that Spacemaker, a secreted protein expressed only in the eyes of insects with open rhabdom systems, acts together with Prominin and the cell adhesion molecule Chaoptin to choreograph the partitioning of rhabdomeres into an open system. Furthermore, the complete loss of *spacemaker* (*spam*) converts an open rhabdom system to a closed one, whereas its targeted expression to photoreceptors of a closed system markedly reorganizes the architecture of the compound eyes to resemble an open system. Our results provide a molecular atlas for the construction of microvillar assemblies and illustrate the critical effect of differences in a single structural protein in morphogenesis.

During ommatidium biogenesis the apical membranes of different photoreceptor cells separate from one another and concomitantly produce the rhabdomeric structure needed for housing the phototransduction machinery. This process involves the precise coordination of membrane–membrane adhesion events both within and between rhabdomeres, and culminates in the production of about 60,000 microvilli per cell, each 1–2 µm in length and 50 nm in diameter^{4,5}. The generation of the inter-rhabdomeral space (IRS; Fig. 1), by which rhabdomeres of the photoreceptor neurons partition from each other, is an essential event in the transition of compound eyes from a closed to an open system. This evolutionary change resulted in a considerable improvement in angular sensitivity, thus allowing the detection of smaller moving objects. We screened adult viable ethylmethane sulphonate-mutagenized *Drosophila* lines⁶ for defects in rhabdomere structure and topology by examining deep-pseudopupil phenotypes^{7,8} and performing an electron-microscopic ultrastructural analysis of pupal and adult eyes in candidate lines. Only 2 of 40 different complementation groups, *spacemaker* (*spam*) and *prominin* (*prom*), had eyes with apparently normal microvilli but showing a striking failure of the rhabdomeres to separate from each other, including a marked loss of IRS, irregular rhabdomere morphology and improper rhabdomeral contacts (Fig. 1a–d). Interestingly, loss-of-function mutations in *spam* completely eliminated the IRS, thus generating ommatidia

resembling the compound eyes of animals with closed rhabdom systems (compare Fig. 1a–c).

To understand how *spam* and *prom* function, both mutants were mapped and the candidate genes were cloned and tested for rescue of the mutant phenotypes by P-element germline transformation (Supplementary Fig. 1c). *spam* encodes a 2,165-amino-acid polypeptide containing several protein motifs commonly found in extracellular molecules: an amino terminus consisting of seven epidermal growth factor (EGF)-like repeats, a linker region containing multiple sites for glycosaminoglycan addition, and a carboxy terminus including four alternating repeats of EGF-like and Laminin G domains (Supplementary Fig. 1a). *prom* encodes a 910-amino-acid Prominin (Prom)-like molecule (Supplementary Fig. 1b), a family of evolutionarily conserved transmembrane proteins often associated with microvilli⁹ but of unknown function.

What is the role of *spam*? Sequence analysis of *spam* indicates that it might function as an extracellular protein (Supplementary Fig. 1). If Spacemaker (Spam) is a secreted molecule, it should localize to the IRS. Indeed, immunocytochemical studies showed that Spam is selectively localized to the IRS (Fig. 1g, h). However, mosaic analysis reveals that even though Spam is capable of diffusing throughout the ommatidium, near wild-type levels of the protein are still required to ensure complete rhabdomere separation (Supplementary Fig. 3b, and data not shown). To determine when *spam* function is required, we placed its transcriptional unit under the control of an inducible heat-shock promoter (UAS-*spacemaker* and heat shock (HS)-GAL4) and subjected mutant animals to temperature shifts at various times during development. Our results demonstrated that *spam* expression at 36–64 h after puparium formation (APF), a window of time coincident with the initiation of rhabdomere biogenesis, is sufficient and necessary for complete rescue of the phenotype (Supplementary Fig. 1c). Late expression (about 72 h APF), using either HS-GAL4 or a late-expressing rhodopsin promoter (Rh1-GAL4), leads only to a partial rescue: Spam is still secreted into the IRS and the apical stalk membranes (the base of the rhabdomeres) are still capable of separating, but the rhabdomeres themselves remain fused in the centre of the ommatidium (Supplementary Fig. 1c). Notably, this phenotype closely resembles that of *prom* mutants (Fig. 1d), suggesting that they might both participate in a common morphogenesis programme.

If *prom* and *spam* function in the same pathway, we would expect genetic interactions between these two loci. Indeed, although both *spam* and *prom* are recessive genes, the *spam prom* double heterozygote shows a failure of the rhabdomeres to separate and the loss of a continuous IRS (Fig. 1e); these results indicate that the two gene products probably act together in the biogenesis of the IRS (see below). To explore the function of Prom, we first examined its spatial and temporal expression pattern. Like Spam, Prom is present at the beginning of rhabdomere biogenesis (48 h APF), when it decorates the entire photoreceptor apical surface (Supplementary Fig. 2a). By the time of eclosion, however, Prom is selectively localized to the stalk membrane and the tips of the microvilli; this is best illustrated by simultaneous labelling with Chaoptin, a photoreceptor adhesion

¹Howard Hughes Medical Institute and Departments of Neurobiology and Neurosciences, University of California at San Diego, La Jolla, California 92093-0649, USA.

protein expressed throughout the perimeter of microvilli (Supplementary Fig. 2c, f, i). Given that Spam is an IRS-secreted molecule and Prom is a rhabdomere integral membrane protein, we examined whether these two proteins interact. When *Drosophila* tissue-culture cells are transfected with *spam*, the protein is secreted into the medium. However, if the same cells are instead transfected together with *prom*, Spam then localizes selectively to the exterior surface of the plasma membrane (Supplementary Fig. 4b). These results indicate that Prom might function as a binding partner for Spam. To demonstrate that Prom in fact recruits extracellular Spam to the plasma membrane, we performed a mixing experiment: RFP-labelled cells expressing *spam* were incubated with green fluorescent protein (GFP)-labelled cells transfected with *prom*. As proposed, secreted Spam now specifically localizes to the surface of Prom-expressing cells (Supplementary Fig. 4f). Taken together, these findings substantiate Prom as a candidate Spam receptor.

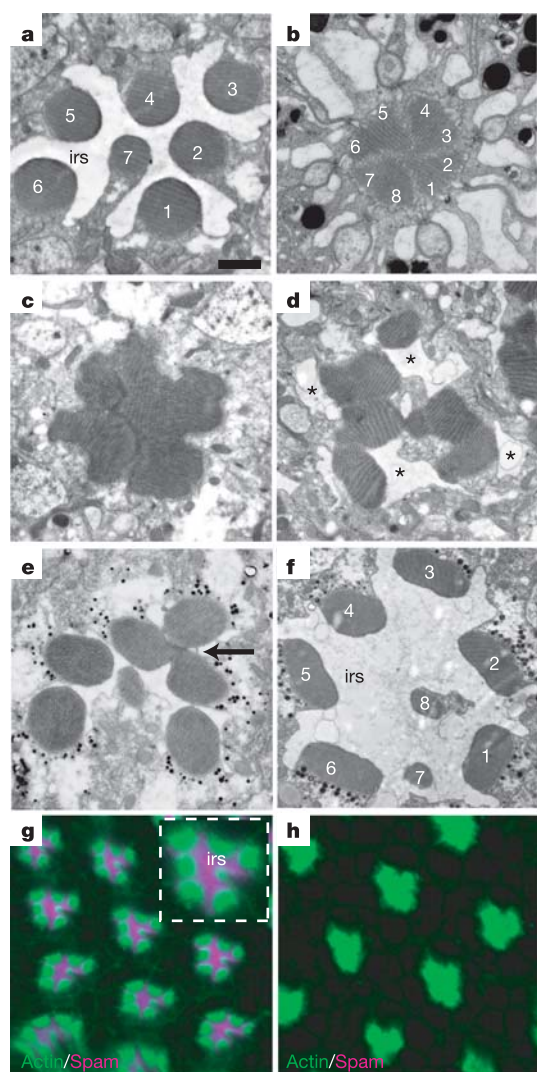


Figure 1 | *spam* and *prom* mutants have severe defects in IRS formation. **a–f**, Electron micrographs showing cross-sections through ommatidia of *Drosophila melanogaster* (**a**), *Apis mellifera* (**b**), *spam*¹ (**c**), *prom*¹ (**d**), *spam*¹ +/+ *prom*¹ (**e**) and flies overexpressing *spam* (GMR-GAL4, UAS-*spam*) (**f**). Asterisks and IRS mark the inter-rhabdomeral space and numbers label each rhabdomere. Note the absence of IRS in *spam* and *prom* mutants, and the incomplete rhabdomeric separation in the transheterozygote (arrow). Scale bar, 1 μ m. **g**, Spam (magenta) expression in wild-type ommatidia; **h**, *spam*¹-null ommatidia. Note Spam localization to the IRS surrounding each of the rhabdomeres (microvilli are labelled green with phalloidin).

What is the function of *prom* and *spam* in orchestrating open rhabdomere development? Microvilli are interconnected by a network of homophilic interactions that ensures their tight and regular packing within rhabdomeres^{10,11}. However, such an arrangement would also render microvilli vulnerable to inter-rhabdomeric adhesion. We suggest that secretion of Spam into the IRS forces the separation of the stalk membranes, pushing the rhabdomeres apart, and that the recruitment of Spam to the microvillar surface by the binding of Prom prevents inter-rhabdomere adhesion. This model is consistent with the phenotypes of *spam* and *prom* mutants, and makes three significant predictions: first, overexpression of *spam* should increase the volume of the IRS by pushing rhabdomeres further away from each other; second, in mosaic ommatidia containing *prominin* mutant cells there should be a significant loss of Spam binding and accumulation surrounding the mutant rhabdomeres; and third, the fused rhabdomeres of *prom* mutants should be rescued by independently reducing or abolishing inter-rhabdomere interactions. As predicted, overexpression of *spam* markedly expands the IRS of wild-type photoreceptors (Fig. 1f), and the presence of Prom in mosaic ommatidia promotes the recruitment of Spam selectively around wild-type rhabdomeres (Supplementary Fig. 3c, d). To alter inter-rhabdomere adhesion, we first needed to determine what molecule was responsible for the inappropriate fusion of rhabdomeres in *prom* and *spam* mutants. We reasoned that Chaoptin, a photoreceptor-specific glycosylphosphatidylinositol-linked membrane protein expressed very early during photoreceptor morphogenesis and required for crosslinking microvilli by means of homophilic interactions, might be a strong candidate^{10–12}. We recognized that the early expression of Chaoptin would ensure the proper packing of microvilli within rhabdomeres and that the subsequent expression of Spam and Prom would guarantee that individual rhabdomeres remain separated in open rhabdom systems. Indeed, removing just one copy of the gene encoding Chaoptin (*chp*) in a *prom* mutant background (*prom*/*prom*; *chp*/+) is sufficient for a

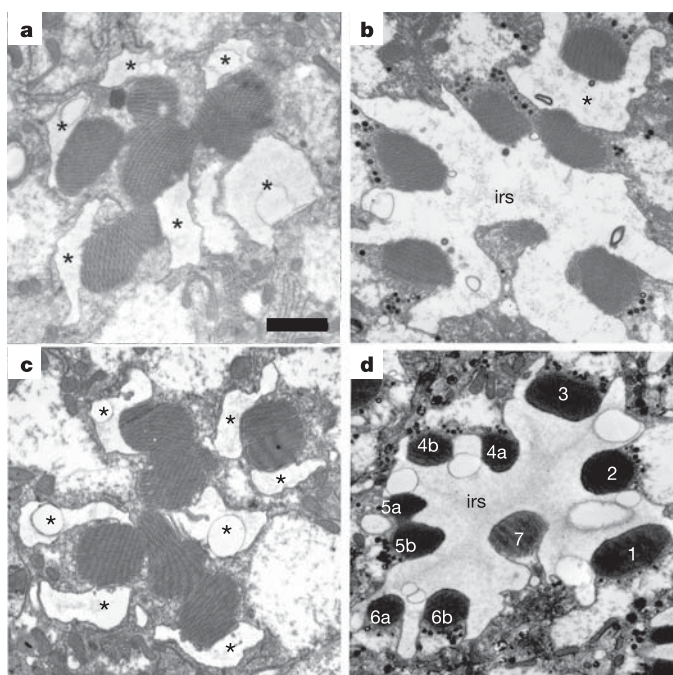


Figure 2 | Prom and Spam cooperate to antagonize the adhesive force of Chaoptin. **a–c**, Electron micrographs showing cross-sections through *prom*/*prom* (**a**), *prom*/*prom*; *chp*/+ (**b**) and *prom* *spam*/*prom* +; *chp*/+ (**c**) ommatidia. **d**, Early expression of Spam produces severely split rhabdomeres. Asterisks denote the IRS. Scale bar, 1 μ m. See Supplementary Fig. 5 for a quantitative analysis of the data.

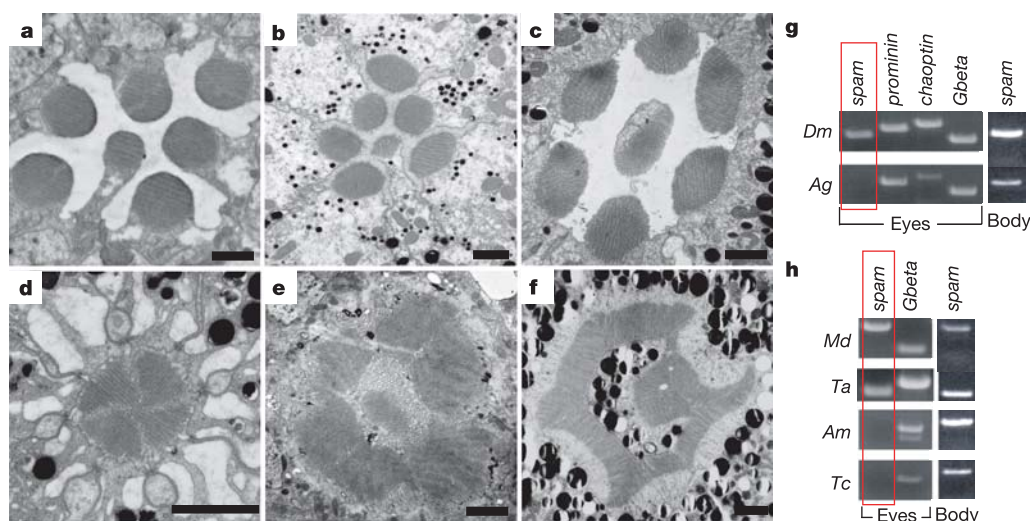


Figure 3 | *spam* is expressed only in insect eyes with an open rhabdomere system. **a–f**, Electron micrographs showing ommatidia of *Drosophila melanogaster* (*Dm*; **a**), *Musca domestica* (*Md*; **b**), *Toxorhynchites amboinensis* (*Ta*; **c**), *Apis mellifera* (*Am*; **d**), *Tribolium castaneum* (*Tc*; **e**) and *Anopheles gambiae* (*Ag*; **f**). Scale bars, 1 μm. **g**, **h**, RT-PCR reactions

examining the expression of *spacemaker*, *prominin*, *chaoptin* and *eye Gbeta* control mRNA in the pupal visual system and bodies of the various species. Multiple *Spam*-specific primers were designed for each species after the cloning of the corresponding homologues.

drastic suppression of the *prom* mutant phenotype (Fig. 2a, b and Supplementary Fig. 5). These data demonstrate that the capacity of *Spam* to push rhabdomeres apart is very sensitive to the amount of adhesive force available (for example *Chaoptin*). Given these results, we anticipated that further reducing the levels of *spam* in this *prom/prom;chp/+* genetic background (that is, bringing the ratio of *spam* to *chp* back to normal levels) should revert the ‘rescued phenotype’ to the mutant state. As proposed, the rhabdomeres are now stuck together (Fig. 2c and Supplementary Fig. 5). Thus, in this genetic background, there is sufficient *Spam* to partition the stalk membranes but not enough to overcome the *Chaoptin*-dependent adhesiveness between adjacent rhabdomeres. A final prediction of these findings is that early expression of *Prom* and *Spam* should result in the development of broken-up rhabdomeres, probably as the *Spam/Prom* complex outcompetes the adhesive force linking microvilli. The data shown in Fig. 2d soundly validate this postulate. Together, these results demonstrate that *Spam*, *Prom* and *Chaoptin* orchestrate the assembly of microvilli, ensure the structural integrity and the partitioning of rhabdomeres, and guarantee the construction of an open rhabdom system.

Are these molecules responsible for the transformation between closed and open rhabdom systems? To examine this possibility we investigated the expression of *Spam*, *Prom* and *Chaoptin* homologues in other insect species containing open or closed rhabdoms (Fig. 3a–f). First, analysis of *Anopheles gambiae* eyes, a closed system, showed that both *Prom* and *Chaoptin* are present but *Spam* is absent from pupal eyes, suggesting that *Spam* might be the critical component for partitioning rhabdomeres (Fig. 3g). To lend further credibility to our hypothesis, we isolated and examined *spam* expression in several other species, two with open rhabdomere systems (the housefly *Musca domestica* and the mosquito *Toxorhynchites amboinensis*) and two with closed ones (the honeybee *Apis mellifera* and the flour beetle *Tribolium castaneum*). As predicted, *Spam* is conspicuously absent from the eyes of insects with closed rhabdomere systems (Fig. 3h), even though *spam* expression is still detected in the body of all of these species, probably reflecting an evolutionarily conserved role¹³ (Fig. 3g). Taken together, these results substantiate *spam* expression as a powerful indicator of visual system architecture and suggest that targeting *spam/prom* function to a closed rhabdom system may well transform it into an open system. To test this final postulate, we focused on the ocellar photoreceptors of *Drosophila*.

Ocelli are simple eyes located at the vertex of the fly head and involved in navigation^{14,15}. They are composed of 70–90 photoreceptor neurons organized in a near-close rhabdomeric arrangement (Fig. 4a, d; note the lack of an IRS). We engineered flies in which

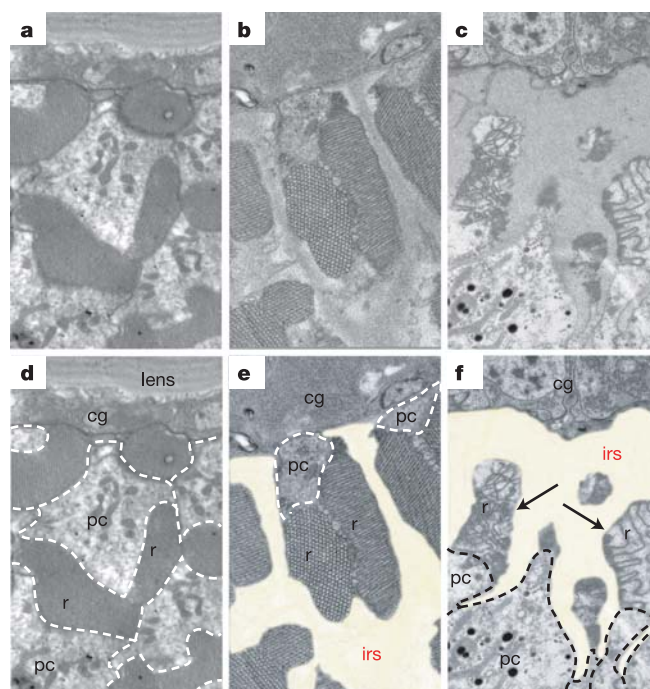


Figure 4 | Expression of *Spam* induces the formation of an IRS. Electron micrographs showing cross-sections through wild-type ocelli (**a**, **d**), ocelli overexpressing *Spam*, *spam/spam;HS-GAL4/UAS-spam* (**b**, **e**) and ocelli strongly overexpressing *Spam*, *GMR-GAL4/UAS-spam* (**c**, **f**). Note the drastic reorganization of the ocellar visual system with increasing amount of *Spam* (see text for details). The bottom panels highlight the IRS (coloured in yellow). Arrows denote the presence of disassembled microvilli within individual rhabdomeres, and the dotted lines outline the body of the photoreceptor cells. Abbreviations: pc, photoreceptor cell; r, rhabdome; cg, corneagenous cells.

Spam expression was targeted to the ocelli under the control of either a weak (HS-GAL4/UAS-*spam* at 29 °C) or a strong (glass multimer reporter (GMR)-GAL4, UAS-*spam*) promoter, and examined eye morphology by electron microscopy. Overexpression of Spam produced a complete reorganization of ocellar eye architecture ranging from the creation of a novel IRS surrounding otherwise structurally intact rhabdomeres (Fig. 4b, e) to the generation of an extensive IRS accompanied by the disintegration of the rhabdomeres into their individual microvilli (GMR-GAL4, UAS-*spam*) (Fig. 4c, f).

Taken together, these studies delineate a vital evolutionary role for Spam in the construction of insect eyes, assign candidate functions to three important proteins and provide a mechanistic path for the assembly of rhabdomeres. Interestingly, human prominin-1 localizes to the base of the outer segments, a subcellular compartment responsible for producing new discs, and frameshift mutations in hProm-1 leads to severe retinal degeneration¹⁶. We suggest that, much as in *Drosophila*, Prom has a key function in the morphogenesis of discs by ensuring their unimpeded folding and stacking into the outer segment.

METHODS

Genetic stocks/experimental animals. The following stocks were used in this study: *cn bw*, Heat Shock-GAL4, RH1-GAL4, GMR-GAL4, *chp*², *prom*¹, and *spam*¹. FRT Stocks: *ey-FLP*;GMR-myrGFP FRT40A, *ey-FLP*;GMR-myrGFP FRT42D, *spam* FRT40A and *prom* FRT42D. P-element-mediated germline transformation and fly manipulations were performed in accordance with standard techniques.

Electron microscopy, immunofluorescence staining, and imaging. Insect heads were fixed in 4% formaldehyde, 3.5% glutaraldehyde, 100 mM cacodylate buffer (pH 7.4), 2 mM CaCl₂ for 3 h at 22 °C, and then fixed again overnight at 4 °C. Heads were rinsed in 100 mM cacodylate buffer and postfixed in 2% buffered osmium for 1 h at 22 °C. The heads were dehydrated through an ethanol series, rinsed with propylene oxide and embedded in EMBED-812 for sectioning. After counterstaining, the sections were viewed and photographed with a Jeol 1200EX II transmission electron microscope. Developing whole retinas were dissected at the appropriate time and processed as described previously¹⁷. Creation and processing of frozen thin sections were performed as described¹⁸. The antibodies used were as follows: Prom, rabbit polyclonal antibody produced by injecting rabbits with the C-terminal peptide ARSKGRHRRSDRSP; Spam, in this study we have shown that the previously unknown antigen for mAb21A6 (ref. 19) is Spam; Chaoptin, mAb2A12 (ref. 19). Rhodamine-fluorescein isothiocyanate (FITC)-conjugated phalloidin (Invitrogen) was used for the detection of F-actin. FITC-conjugated, Cy5-conjugated and Red-X-conjugated secondary antibodies were obtained from Jackson ImmunoResearch Laboratories. All images were captured on a Bio-Rad MRC1024 or Zeiss LSM510 confocal microscope and imported into Adobe Photoshop. Antibodies were tested and validated for specificity and selectivity by performing staining reactions in control and loss-of-function mutant alleles. The anti-Spam, anti-Prominin and anti-Chaoptin¹¹ staining is completely abolished in null alleles (see Fig. 1h, Supplementary Table 1 and Supplementary Fig. 3a).

Transgenic constructs. Complementary DNAs representing *spacemaker* and *prominin* were cloned into pUAST²⁰ and transformed into *Drosophila*. The rescue of *spacemaker* and *prominin* was assayed in *spam*/*spam*;HS-GAL4/UAS-*spam*, *spam*/*spam*;RH1-GAL4/UAS-*spam*, *prom*/*prom*;HS-GAL4/UAS-*prom*, and *prom*/*prom*;RH1-GAL4/UAS-*prom* and reared at 22 °C. For the temperature-shift experiments white puparia were collected and transferred to 29 °C at various times of development and assayed by electron microscopy for rescue. For overexpression of *Spacemaker* HS-GAL4/UAS-*spam* larvae were collected, heat shocked for 1 h at 37 °C every 24 h and reared at 29 °C until eclosion.

Transfection assays. *Drosophila* Schneider 2 cells were transfected with Effectene (Qiagen), and 48 h after transfection they were plated on polylysine-coated slides for fixing and imaging. Cells were transfected with pUAST-Spacemaker, pUAST-Prominin, pUAST-GFP and pUAST-mRFP and to drive expression they were co-transfected with pTub-GAL4 (ref. 21). All fixations and staining were performed in the absence of detergent to detect the surface localization of Spam.

RNA isolation and RT-PCR. Extraction of total RNA from insect tissues was performed with Trizol (Invitrogen). In addition, all RNA samples were treated

with DNase and first-strand synthesis was primed with oligo(dT) and completed with Superscript First Strand Synthesis (Invitrogen). Species-specific primer sets were used for PCR amplification (see Supplementary Table 2).

Received 9 May; accepted 28 July 2006.

Published online 1 October 2006.

- Land, M. F. in *Facets of Vision* (eds Stravenga, D. G. & Hardie, R. C.) 90–111 (Springer, Berlin, 1989).
- Nilsson, D. E. in *Facets of Vision* (eds Stravenga, D. G. & Hardie, R. C.) 30–73 (Springer, Berlin, 1989).
- Shaw, S. R. in *Facets of Vision* (eds Stravenga, D. G. & Hardie, R. C.) 186–212 (Springer, Berlin, 1989).
- Leonard, D. S., Bowman, V. D., Ready, D. F. & Pak, W. L. Degeneration of photoreceptors in rhodopsin mutants of *Drosophila*. *J. Neurobiol.* **23**, 605–626 (1992).
- Kumar, J. P. & Ready, D. F. Rhodopsin plays an essential structural role in *Drosophila* photoreceptor development. *Development* **121**, 4359–4370 (1995).
- Koundakjian, E. J., Cowan, D. M., Hardy, R. W. & Becker, A. H. The Zuker collection: a resource for the analysis of autosomal gene function in *Drosophila melanogaster*. *Genetics* **167**, 203–206 (2004).
- Franceschini, N. & Kirschfeld, K. Pseudopupil phenomena in the *Drosophila* compound eye. *Kybernetik* **9**, 159–182 (1971).
- Franceschini, N. in *Information Processing in the Visual Systems of Arthropods* (ed. Wehner, R.) 75–82 (Springer, Berlin, 1972).
- Corbeil, D., Roper, K., Fargeas, C. A., Joester, A. & Huttner, W. B. Prominin: a story of cholesterol, plasma membrane protrusions and human pathology. *Traffic* **2**, 82–91 (2001).
- Krantz, D. E. & Zipursky, S. L. *Drosophila* chaoptin, a member of the leucine-rich repeat family, is a photoreceptor cell-specific adhesion molecule. *EMBO J.* **9**, 1969–1977 (1990).
- Van Vactor, D. Jr, Krantz, D. E., Reinke, R. & Zipursky, S. L. Analysis of mutants in chaoptin, a photoreceptor cell-specific glycoprotein in *Drosophila*, reveals its role in cellular morphogenesis. *Cell* **52**, 281–290 (1988).
- Reinke, R., Krantz, D. E., Yen, D. & Zipursky, S. L. Chaoptin, a cell surface glycoprotein required for *Drosophila* photoreceptor cell morphogenesis, contains a repeat motif found in yeast and human. *Cell* **52**, 291–301 (1988).
- Avidor-Reiss, T. et al. Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* **117**, 527–539 (2004).
- Pollock, J. A. & Benzer, S. Transcript localization of four opsin genes in the three visual organs of *Drosophila*; RH2 is ocellus specific. *Nature* **333**, 779–782 (1988).
- Yoon, C. S., Hirose, K. & Suzuki, E. Studies on the structure of ocellar photoreceptor cells of *Drosophila melanogaster* with special reference to subrhabdomeric cisternae. *Cell Tissue Res.* **284**, 77–85 (1996).
- Maw, M. A. et al. A frameshift mutation in prominin (mouse)-like 1 causes human retinal degeneration. *Hum. Mol. Genet.* **9**, 27–34 (2000).
- Zelhof, A. C. & Hardy, R. W. WASp is required for the correct temporal morphogenesis of rhabdomere microvilli. *J. Cell Biol.* **164**, 417–426 (2004).
- Tsunoda, S., Sun, Y., Suzuki, E. & Zuker, C. Independent anchoring and assembly mechanisms of INAD signaling complexes in *Drosophila* photoreceptors. *J. Neurosci.* **21**, 150–158 (2001).
- Zipursky, S. L., Venkatesh, T. R., Teplow, D. B. & Benzer, S. Neuronal development in the *Drosophila* retina: monoclonal antibodies as molecular probes. *Cell* **36**, 15–26 (1984).
- Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
- Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **22**, 451–461 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Meerloo for help with cryosections; D. Cowan for help with P-element injections; Glenn Apiaries, L. McCuiston, R. Denell, S. Brown, T. Shippy, M. Lorenzen, P. Atkinson and A. James for providing various insect species; and W. McGinnis, N. Ryba, S. L. Zipursky, K. Kirschfeld and members of the Zuker laboratory for comments. This work was supported in part by a grant from the National Eye Institute. C.S.Z. is an investigator of the Howard Hughes Medical Institute.

Author Information The cDNA sequences for *spacemaker* and *prominin* are deposited in GenBank under accession numbers DQ780942 and DQ780943, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.S.Z. (charles@flyeye.ucsd.edu) or A.C.Z. (azelhof@flyeye.ucsd.edu).

LETTERS

Bidirectional control of CNS capillary diameter by pericytes

Claire M. Peppiatt^{1*}, Clare Howarth^{1*}, Peter Mobbs¹ & David Attwell¹

Neural activity increases local blood flow in the central nervous system (CNS), which is the basis of BOLD (blood oxygen level dependent) and PET (positron emission tomography) functional imaging techniques^{1–3}. Blood flow is assumed to be regulated by precapillary arterioles, because capillaries lack smooth muscle. However, most (65%) noradrenergic innervation of CNS blood vessels terminates near capillaries rather than arterioles⁴, and in muscle and brain a dilatory signal propagates from vessels near metabolically active cells to precapillary arterioles^{5,6}, suggesting that blood flow control is initiated in capillaries. Pericytes, which are apposed to CNS capillaries and contain contractile proteins⁷, could initiate such signalling. Here we show that pericytes can control capillary diameter in whole retina and cerebellar slices. Electrical stimulation of retinal pericytes evoked a localized capillary constriction, which propagated at $\sim 2 \mu\text{m s}^{-1}$ to constrict distant pericytes. Superfused ATP in retina or noradrenaline in cerebellum resulted in constriction of capillaries by pericytes, and glutamate reversed the constriction produced by noradrenaline. Electrical stimulation or puffing GABA (γ -amino butyric acid) receptor blockers in the inner retina also evoked pericyte constriction. In simulated ischaemia, some pericytes constricted capillaries. Pericytes are probably modulators of blood flow in response to changes in neural activity, which may contribute to functional imaging signals and to CNS vascular disease.

Neurotransmitters cause cultured pericytes to contract⁸, and raise intracellular Ca^{2+} concentration ($[\text{Ca}^{2+}]_i$) and constrict pericytes on isolated retinal microvessels^{9–11}. Because neocortical pericytes *in situ* show neurally evoked increases in $[\text{Ca}^{2+}]_i$ (ref. 12), pericytes might control blood flow downstream of arterioles (Fig. 1a). It is unknown, however, whether pericytes regulate capillary diameter *in situ* in retina or brain.

In retina and cerebellum, an antibody to the NG2 proteoglycan¹³ identified two pericyte classes: those on straight parts of capillaries, separated by $34.2 \pm 3.6 \mu\text{m}$ (mean $\pm$ s.e.m.; $n = 24$) in retina; and those at capillary junctions (Fig. 1a–c). Pericytes have processes along and around the capillary (Fig. 1d). Loading dye into retinal pericytes by whole-cell clamping revealed processes similar to claws around the capillary (Fig. 1e), extending $11.3 \pm 0.7 \mu\text{m}$ ($n = 3$) from the soma, providing an anatomical substrate for generating the capillary constriction described below.

We stimulated retinal pericytes electrically, with a pipette pressed on their soma, aiming to raise $[\text{Ca}^{2+}]_i$ because in isolated vessels pericyte constriction correlates with a rise in $[\text{Ca}^{2+}]_i$ (ref. 11). Stimulation constricted most (90% of 30) pericytes studied (Fig. 2a–d and Supplementary Movie 1), starting at $3.6 \pm 1.6 \text{ s}$ and reaching half the maximum constriction at $12.6 \pm 3.6 \text{ s}$ after initiating stimulation. After stimulation, constriction relaxed 50% in $60.3 \pm 12.4 \text{ s}$. For 17 pericytes on straight parts of capillaries (initial diameter $8.6 \pm 0.5 \mu\text{m}$), the mean vessel constriction was $73 \pm 5\%$

(Fig. 2e). Stimulation constricted pericytes at branch points (initial diameter $5.8 \pm 0.6 \mu\text{m}$) by $76 \pm 8\%$ ($n = 11$). Stimulating the capillary wall between pericytes did not constrict the capillary or adjacent pericytes ($n = 7$; Fig. 2e); thus, endothelial cells and astrocyte endfeet do not constrict.

Removing extracellular Ca^{2+} evoked a dilation of capillaries near pericytes (by 28% in four vessels, $P = 0.023$) that slowly decayed, but evoked no dilation at non-pericyte sites ($P = 0.22$; Fig. 2f, g). Removing Ca^{2+} abolished the stimulation-evoked constriction (Fig. 2f, h), consistent with constriction being triggered by a rise in $[\text{Ca}^{2+}]_i$. When pericytes were stimulated to constrict, sometimes distant pericytes also constricted subsequently (Fig. 3a and Supplementary Movie 1); however, the capillary between the pericytes did not constrict (Fig. 3b). The constriction of distant pericytes was not due to extracellular stimulus spread because stimulating between pericytes did not evoke a constriction (Fig. 2e). The speed of the constriction spread was $2.0 \pm 1.3 \mu\text{m s}^{-1}$ ($n = 4$).

In brain, $[\text{Ca}^{2+}]_i$ increases in astrocytes regulate arteriole diameter^{14–16}. ATP is released when $[\text{Ca}^{2+}]_i$ waves spread between retinal glial cells¹⁷, and constricts pericytes on isolated vessels by raising $[\text{Ca}^{2+}]_i$ through P2X₇ and P2Y receptors⁹. In the intact retina, ATP and the P2Y receptor agonist UTP constricted capillaries near pericytes, leaving non-pericyte regions unconstricted (Fig. 3c–e and Supplementary Movie 2). This occurred in pericytes on straight parts of capillaries and at branch points, but not in all pericytes: 50–100 μM UTP constricted vessels at 30% (12 of 40) of pericytes, whereas 0.5–1 mM ATP constricted vessels at 25% of pericytes (5 of 20; insignificantly different, $P = 0.92$). Peak constrictions evoked by 100 μM UTP ($56 \pm 9\%$, $n = 10$, initial diameter $7.1 \pm 0.9 \mu\text{m}$) and 1 mM ATP ($35 \pm 6\%$, $n = 4$, initial diameter $7.9 \pm 2.3 \mu\text{m}$) were not significantly different ($P = 0.19$). The lack of response of some pericytes is similar to findings on isolated vessels (a P2X₇ agonist contracted 37% of pericytes and capillary constriction occurred in only a third of these or 12% of the total⁹, and a muscarinic agonist contracted only 10% of pericytes¹¹). UTP- and ATP-evoked constrictions showed desensitization (Fig. 3d) to $34 \pm 3\%$ and $32 \pm 17\%$ of their initial value. Glutamate (500 μM) did not affect retinal capillary diameter ($n = 39$), and noradrenaline (0.3–10 μM) constricted pericytes on only 3 of 61 retinal capillaries despite always constricting arterioles (Supplementary Fig. 1).

We tested whether ATP release from underlying glia might explain the stimulation-evoked pericyte constriction in Fig. 2, or the propagating constriction in Fig. 3a. Applying P2 receptor blockers (suramin, 100 μM ; PPADS, 100 μM), which inhibit the spread of ATP-releasing Ca^{2+} waves through retinal glia¹⁷ and would reduce the actions of any released ATP on the P2Y₂, P2Y₄ and P2X₇ receptors that trigger pericyte contraction⁹, did not affect the fraction of pericytes responding to electrical stimulation or the constriction seen ($P = 0.96$ and 0.63, respectively; $n = 10$), eliminating the possibility that pericyte

¹Department of Physiology, University College London, Gower Street, London WC1E 6BT, UK.

*These authors contributed equally to this work.

stimulation releases ATP from glia that constricts pericytes. Furthermore, although pericyte stimulation sometimes evoked a Ca^{2+} wave (see Methods) in underlying glia, as did direct glial stimulation, the wave propagation speed ($10.5 \pm 1.2 \mu\text{m s}^{-1}$, $n = 8$; three stimulating pericytes, five stimulating glia, no significant difference between pericytes and glia) was fivefold faster ($P = 0.0014$) than the propagation of the constriction described above. This suggests that the spread of constriction is not produced by a glial Ca^{2+} wave but by a different mechanism, possibly an electrical signal in endothelial cells or an effect of the pressure change generated by constriction of the stimulated pericyte. Pericyte-pericyte communication could enable the detection of neural activity near capillaries to be propagated back

to arterioles to amplify blood flow changes⁶, as in muscle⁵.

To assess whether neuronal activity can regulate capillary diameter through pericytes, we stimulated electrically or puffed solution containing blockers of GABA_A and GABA_C receptors (GABAzine, $100 \mu\text{M}$; TPMPA, $100 \mu\text{M}$) within the retina near the inner plexiform layer. Stimulation (Fig. 3f, g) evoked a pericyte constriction of $39 \pm 6\%$ (initial diameter $4.6 \pm 0.4 \mu\text{m}$) in 13 of 43 capillaries tested, which was inhibited by the action potential blocker TTX ($P = 0.034$; a second stimulus evoked no constriction in six of six cells in $1 \mu\text{M}$ TTX, but constricted four of five cells without TTX). GABA blockers evoked a constriction of $69 \pm 12\%$ (initial diameter $4.8 \pm 0.4 \mu\text{m}$, latency $34 \pm 3 \text{ s}$, in 4 of 41 pericytes; Fig. 3h, i). The effect of GABA blockers suggests that there is endogenously active neurotransmitter signalling to pericytes.

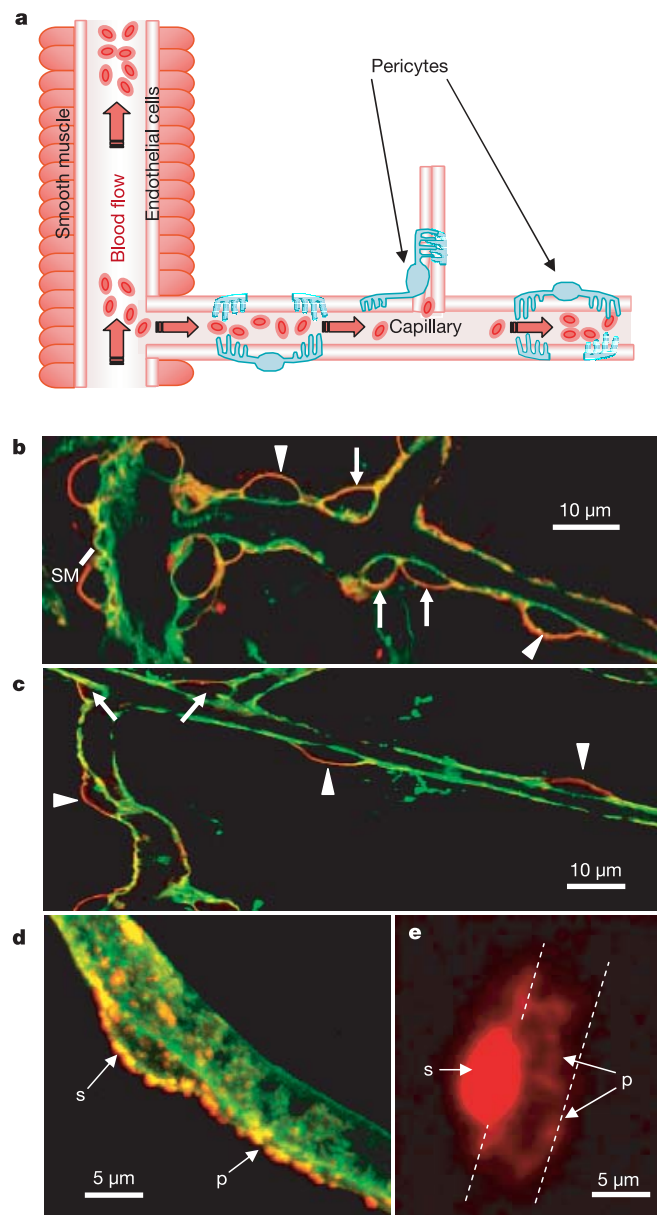


Figure 1 | Pericyte anatomy confers flow-regulating capability downstream of arterioles. **a**, Potential blood flow control sites in cerebral vasculature: arteriolar smooth muscle, and pericytes on capillaries. **b**, Cerebellar molecular layer arteriole (left), surrounded by smooth muscle (SM), giving off a capillary. Capillaries are labelled with isolectin B4 (green); pericytes labelled with NG2 antibody (red) are located on the straight part of capillaries (arrowheads) and at junctions (arrows). **c**, Retinal capillaries. **d**, Soma (s) of cerebellar pericyte gives off processes (p) running along and around capillary. **e**, Dye fill of retinal pericyte reveals processes running around capillary (broken lines).

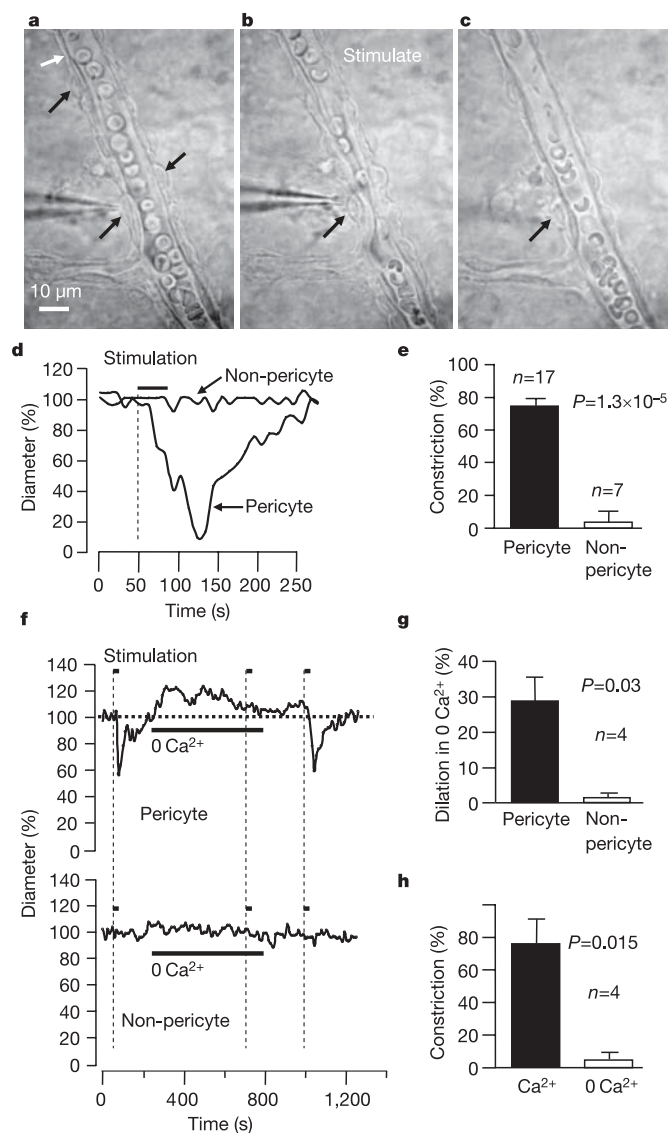


Figure 2 | Electrical stimulation evokes Ca^{2+} -dependent localized constriction of retinal pericytes. **a–c**, Capillary with pericytes (black arrows) before (**a**), during (**b**) and after (**c**) stimulation. Erythrocytes are present in the capillary; thin structures outside the capillary are astrocyte endfeet. **d**, Diameter of capillary in **a–c** at stimulated pericyte and a non-pericyte site (white arrow in **a**). **e**, Mean constriction when stimulating at pericyte or non-pericyte sites. **f**, Effect of removing extracellular Ca^{2+} on resting diameter and response to pericyte stimulation, at pericyte and nearby non-pericyte sites. **g**, Mean dilation produced by removing Ca^{2+} . **h**, Effect of Ca^{2+} removal on pericyte constriction, as in **f**. Error bars in **e**, **g** and **h** indicate the s.e.m.

In cerebellar slices, noradrenaline (1–2.5 μM) constricted 50% (10 of 20) of capillaries at spatially restricted locations near pericytes (Fig. 4a and Supplementary Movie 3). In constricting capillaries, noradrenaline reduced capillary diameter by $63 \pm 8\%$ (initial diameter $7.2 \pm 0.8 \mu\text{m}$) at pericytes, but did not affect the diameter at non-pericyte locations (Fig. 4b, c). In all seven constricting vessels tested, superimposing glutamate (100–500 μM) on noradrenaline decreased

the noradrenaline-evoked constriction—that is, it dilated the vessel—by $40 \pm 8\%$ of the precontracted diameter at the pericyte, but had no effect elsewhere (Fig. 4b, d, and Supplementary Movie 3; 100 μM and 500 μM glutamate produced a $55 \pm 10\%$, and a $34 \pm 9\%$ dilation in two and five vessels; not significantly different, $P = 0.2$). In one capillary, glutamate dilated the vessel beyond the initial diameter that it had in the absence of noradrenaline (by 14%),

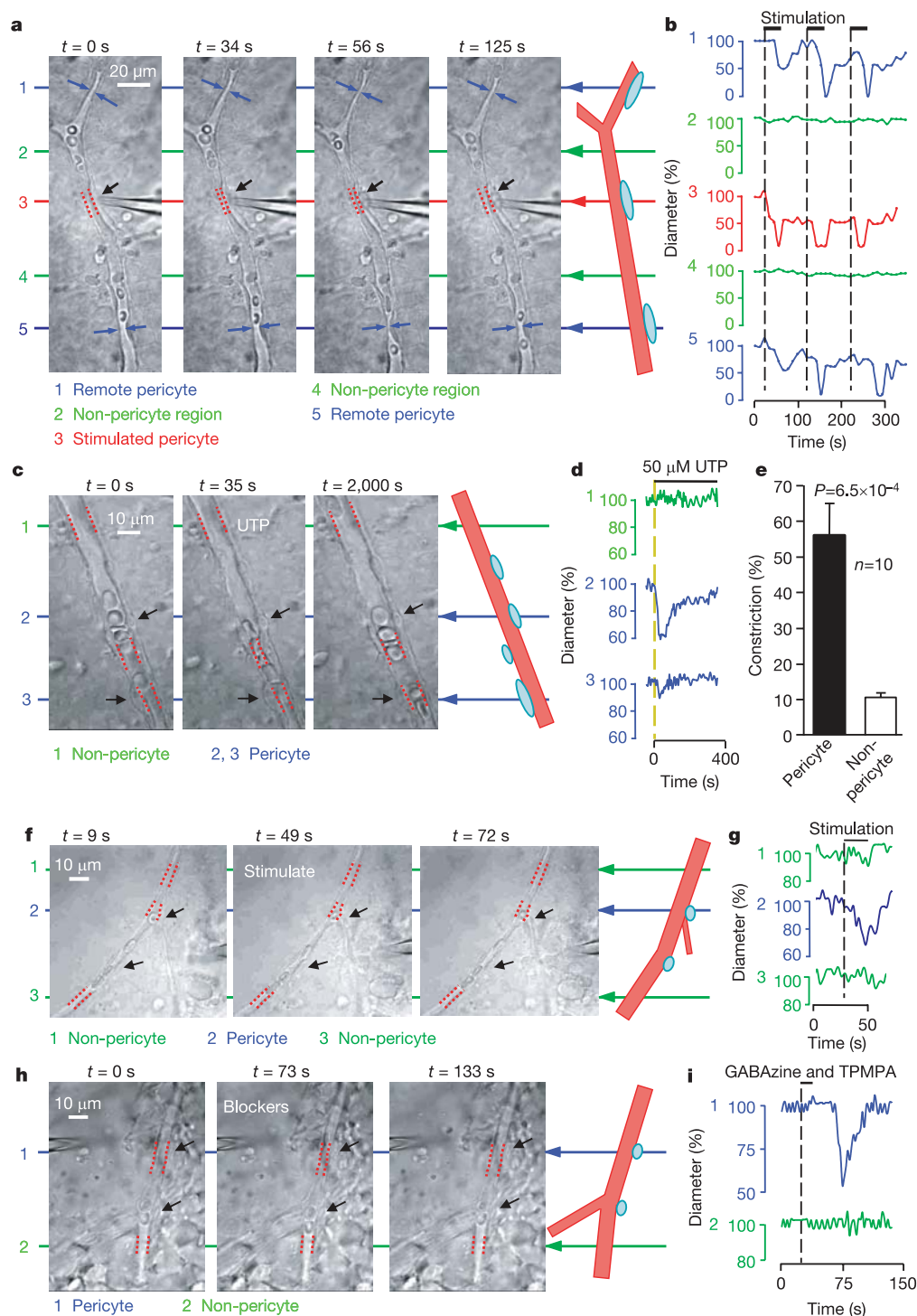


Figure 3 | Propagation and transmitter evocation of retinal pericyte constriction. **a**, Electrically stimulating a pericyte (black arrow) evokes local constriction (red dashes indicate vessel diameter), followed by constriction of distant pericytes (blue arrows). **b**, No constriction of intervening non-pericyte regions (green). **c**, **d**, UTP constricts two pericytes (black arrows), but not non-pericyte regions. **e**, Mean UTP-evoked constriction. Error bars

indicate the s.e.m. **f**, **g**, Pericyte constriction (top black arrows; lower black arrows indicate another pericyte) evoked by electrical stimulation (electrode, right) near the inner plexiform layer. **h**, **i**, Constriction of pericyte (top black arrows; lower black arrows indicate another pericyte) evoked by puffing GABA receptor blockers (electrode, top left) near the inner plexiform layer.

and in 4 (of 20) vessels tested, 500 μM glutamate alone (without noradrenaline) dilated capillaries (by $22 \pm 9\%$), showing that glutamate can produce a net dilation. Thus, pericytes can control capillary diameter in both the brain and the retina.

After ischaemia, cerebral blood flow is initially increased, but later reduced owing to defective vasodilatory pathways^{18–20}. We simulated retinal ischaemia, which leads to a regenerative ‘anoxic depolarization’²¹ associated with increased opacity. This opacity occurred after 435 ± 24 s. Before the anoxic depolarization, some pericytes (4/38) constricted capillaries (Fig. 4e, f) by $53 \pm 9\%$ (initial diameter $6.8 \pm 0.4 \mu\text{m}$).

Our data establish pericytes as likely regulators of blood flow at the capillary level. First, we have shown that neurotransmitters evoke pericyte-mediated capillary constriction *in situ*. This had been previously shown only in isolated retinal capillaries, and it was claimed that vasoactive agents do not affect capillary diameter *in situ*²² (but see ref. 23). Second, we have shown that capillary diameter changes are generated by pericytes, not endothelial cells, because the changes do not occur in pericyte-free regions of capillaries. Third, although

retinal vessels have more pericytes than brain capillaries²⁴, noradrenaline-evoked constriction of cerebellar capillaries indicates that pericytes can regulate blood flow throughout the brain. Fourth, we have shown that decreasing pericyte contractile tone dilates capillaries: Ca^{2+} removal dilates retinal capillaries and glutamate dilates cerebellar capillaries through pericytes. Glutamate may release nitric oxide, which inhibits Ca^{2+} influx in retinal pericytes and decreases the contraction of cultured pericytes^{25,26}. This could contribute to the increase of blood flow evoked by neural activity. Last, we have shown that pericytes may contribute to the vascular response to ischaemia.

The constriction that ATP and noradrenaline produce (35% in retina and 63% in cerebellum, respectively) would increase flow resistance 5.6-fold and 53-fold by Poiseuille’s law, and might prevent erythrocytes passing through capillaries. Thus, pericytes could significantly redirect blood flow at the capillary level. It is unclear whether ATP and noradrenaline act directly on pericytes⁹ or through astrocytes^{27,28}.

The low fraction of pericytes responding to neurotransmitters

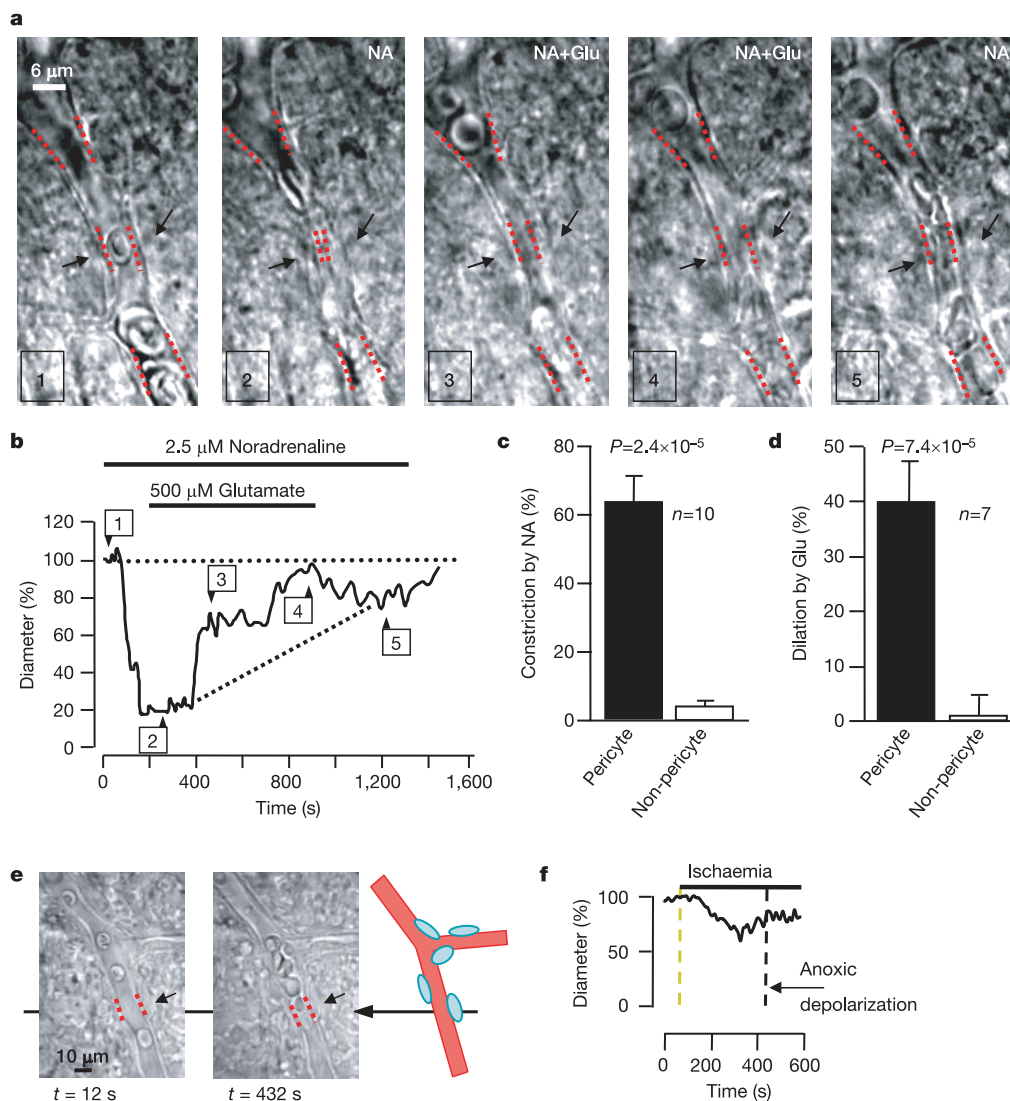


Figure 4 | Effects of neurotransmitters on cerebellar molecular layer capillary and effects of ischaemia on retinal capillary. **a**, Localized noradrenaline-evoked constriction near pericytes (black arrows, clearest in panel 4), and dilation by superimposed glutamate. Noradrenaline-evoked constriction does not reflect fluid movement caused by arterioles constricting, because this would make the capillary dilate. NA, noradrenaline;

Glu, glutamate. **b**, Diameter at pericyte; numbers correspond to the images in **a**. **c**, Mean constriction by noradrenaline at pericyte and non-pericyte sites. **d**, Mean dilation (reduction of noradrenaline-evoked constriction) by glutamate (as percentage of pre-noradrenaline diameter). Errors bars in **c** and **d** indicate the s.e.m. **e**, Retinal capillary before ischaemia and during ischaemia before anoxic depolarization. **f**, Diameter at pericyte in **e**.

(25% to ATP in retina, 50% to noradrenaline in cerebellum) contrasts with the reliability of electrical-stimulation-evoked constriction (90%). This may reflect variable expression of transmitter receptors on pericytes at different vascular sites, or a loss of factors released by blood flow. The lack of blood flow and pressure in brain slices may also slow the capillary responses reported here: arteriolar dilations produced by rises in $[Ca^{2+}]_i$ in astrocytes are faster *in vivo* than in slices^{14,16}. In addition, although some pericytes do not visibly constrict, the stiffness of their processes around the capillary may be increased by ATP or noradrenaline, opposing dilation by other agents or by the deformation necessary for erythrocytes to pass along small capillaries.

These data suggest that pericytes may contribute to regulating cerebral blood flow in health and disease. Spatially restricted constrictions of brain microvessels attributed to arteriole smooth muscle^{29,30} may in fact have been mediated by pericytes. These results challenge the idea that arterioles are solely responsible for the blood flow increase evoked by neural activity that underlies functional imaging techniques (see Supplementary Information).

METHODS

Preparations. Isolated rat retina (vitreous side up), or 200 μ m cerebellar slices, from rats at postnatal day 14–21 (P14–P21), were superfused with solution (33–35 °C) containing (in mM): NaCl, 124; NaHCO₃, 26; NaH₂PO₄, 1; KCl, 2.5; CaCl₂, 1.8; MgCl₂, 2; D-glucose, 10 (bubbled with 95% O₂/5% CO₂; pH 7.4). Kynurenic acid (1 mM, blocks glutamate receptors) was included in the dissection and tissue storage solutions. Ischaemia was simulated by replacing 10 mM glucose with 7 mM sucrose, bubbling 95% N₂/5% CO₂, and adding glycolysis and oxidative phosphorylation blockers (2 mM sodium iodoacetate, 25 μ M antimycin).

Imaging. Capillaries, lacking the continuous smooth muscle around arterioles (Supplementary Fig. 1) and <11 μ m diameter, were imaged every 2–10 s. Capillaries, unlike arterioles (Supplementary Fig. 1), showed pericyte-mediated, spatially restricted constrictions in response to transmitters. For Ca²⁺ imaging, retinas were incubated with Fluo-4-AM (70 min, room temperature); fluorescence was excited at 475 nm and measured at 535 nm; surface glia were stimulated electrically. Pericytes were labelled with NG2 antibody (Chemicon), which in P21 and older tissue labels mainly pericyte somata¹³, using Alexa-555-conjugated goat anti-rabbit secondary antibody. Blood vessels were labelled with Alexa-488-conjugated isolectin B₄ (Invitrogen).

Electrophysiology. Retinas were pre-treated with collagenase (1 mg ml⁻¹, 30 min, 37 °C) to remove connective tissue. Pericytes were then whole-cell clamped with ~8-M Ω electrodes containing Alexa488 (1 mg ml⁻¹) and (in mM): potassium gluconate, 130; NaCl, 10; MgCl₂, 1; CaCl₂, 0.1; HEPES, 10; K₂-EGTA, 1.1; Na₂-ATP, 1 (pH 7.0 with KOH). Similar electrodes filled with extracellular solution were used either to stimulate pericytes, by pressing the electrode against the cell and applying voltage pulses (40–90 V, 0.02–0.2 ms, 9 Hz; three pulses evoked constriction); or to stimulate inner retinal neurons, by placing the electrode tip 22 μ m below the retinal surface and applying voltage pulses (60–90 V, 0.02–0.06 ms, 9–20 Hz for 20–30 s; constriction could not be due to extracellular current spread directly activating pericytes because no such response was seen when stimulation was applied to the capillary wall between pericytes).

Statistics. Data are the mean $\pm$ s.e.m. *P* values are from two-tailed Student's *t*-tests or χ^2 -tests.

Received 15 May; accepted 24 August 2006.

Published online 1 October 2006.

- Roy, C. & Sherrington, C. On the regulation of the blood supply of the brain. *J. Physiol. (Lond.)* **11**, 85–100 (1890).
- Raichle, M. E. Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proc. Natl Acad. Sci. USA* **95**, 765–772 (1998).
- Attwell, D. & Iadecola, C. The neural basis of functional brain imaging signals. *Trends Neurosci.* **25**, 621–625 (2002).
- Cohen, Z., Molinatti, G. & Hamel, E. Astroglial and vascular interactions of noradrenaline terminals in the rat cerebral cortex. *J. Cereb. Blood Flow Metab.* **17**, 894–904 (1997).

- Berg, B. R., Cohen, K. D. & Sarelius, I. H. Direct coupling between blood flow and metabolism at the capillary level in striated muscle. *Am. J. Physiol.* **272**, H2693–H2700 (1997).
- Iadecola, C., Yang, G., Ebner, T. J. & Chen, G. Local and propagated vascular responses evoked by focal synaptic activity in cerebellar cortex. *J. Neurophysiol.* **78**, 651–659 (1997).
- Herman, I. M. & D'Amore, P. A. Microvascular pericytes contain muscle and nonmuscle actins. *J. Cell Biol.* **101**, 43–52 (1985).
- Rucker, H. K., Wynder, H. J. & Thomas, W. E. Cellular mechanisms of CNS pericytes. *Brain Res. Bull.* **51**, 363–369 (2000).
- Kawamura, H. et al. ATP: a vasoactive signal in the pericyte-containing microvasculature of the rat retina. *J. Physiol. (Lond.)* **551**, 787–799 (2003).
- Kawamura, H. et al. Effects of angiotensin II on the pericyte-containing microvasculature of the rat retina. *J. Physiol. (Lond.)* **561**, 671–683 (2004).
- Wu, D. M., Kawamura, H., Sakagami, K., Kobayashi, M. & Puro, D. G. Cholinergic regulation of pericyte-containing retinal microvessels. *Am. J. Physiol. Heart Circ. Physiol.* **284**, H2083–H2090 (2003).
- Hirase, H., Creso, J., Singleton, M., Bartho, P. & Buzsaki, G. Two-photon imaging of brain pericytes *in vivo* using dextran-conjugated dyes. *Glia* **46**, 95–100 (2004).
- Hughes, S. & Chan-Ling, T. Characterization of smooth muscle cell and pericyte differentiation in the rat retina *in vivo*. *Invest. Ophthalmol. Vis. Sci.* **45**, 2795–2806 (2004).
- Zonta, M. et al. Neuron-to-astrocyte signaling is central to the dynamic control of brain microcirculation. *Nature Neurosci.* **6**, 43–50 (2003).
- Mulligan, S. J. & MacVicar, B. A. Calcium transients in astrocyte endfeet cause cerebrovascular constrictions. *Nature* **431**, 195–199 (2004).
- Takano, T. et al. Astrocyte-mediated control of cerebral blood flow. *Nature Neurosci.* **9**, 260–267 (2006).
- Newman, E. A. Propagation of intercellular calcium waves in retinal astrocytes and Muller cells. *J. Neurosci.* **21**, 2215–2223 (2001).
- Leffler, C. W., Beasley, D. G. & Busija, D. W. Cerebral ischemia alters cerebral microvascular reactivity in newborn pigs. *Am. J. Physiol.* **257**, H266–H271 (1989).
- Nelson, C. W., Wei, E. P., Povlishock, J. T., Kontos, H. A. & Moskowitz, M. A. Oxygen radicals in cerebral ischemia. *Am. J. Physiol.* **263**, H1356–H1362 (1992).
- Hauck, E. F., Apostel, S., Hoffmann, J. F., Heimann, A. & Kempinski, O. Capillary flow and diameter changes during reperfusion after global cerebral ischemia studied by intravital video microscopy. *J. Cereb. Blood Flow Metab.* **24**, 383–391 (2004).
- Hansen, A. J. Effect of anoxia on ion distribution in the brain. *Physiol. Rev.* **65**, 101–148 (1985).
- Butryn, R. K., Ruan, H., Hull, C. M. & Frank, R. N. Vasoactive agonists do not change the caliber of retinal capillaries of the rat. *Microvasc. Res.* **50**, 80–93 (1995).
- Schonfelder, U., Hofer, A., Paul, M. & Funk, R. H. *In situ* observation of living pericytes in rat retinal capillaries. *Microvasc. Res.* **56**, 22–29 (1998).
- Shepro, D. & Morel, N. M. Pericyte physiology. *FASEB J.* **7**, 1031–1038 (1993).
- Sakagami, K., Kawamura, H., Wu, D. M. & Puro, D. G. Nitric oxide/cGMP-induced inhibition of calcium and chloride currents in retinal pericytes. *Microvasc. Res.* **62**, 196–203 (2001).
- Haefliger, I. O., Zschauer, A. & Anderson, D. R. Relaxation of retinal pericyte contractile tone through the nitric oxide–cyclic guanosine monophosphate pathway. *Invest. Ophthalmol.* **35**, 991–997 (1994).
- Simard, M., Arcuino, G., Takano, T., Liu, O. S. & Nedergaard, M. Signaling at the gliovascular interface. *J. Neurosci.* **23**, 9254–9262 (2003).
- Duffy, S. & MacVicar, B. A. Adrenergic calcium signaling in astrocyte networks within the hippocampal slice. *J. Neurosci.* **15**, 5535–5550 (1995).
- Cauli, B. et al. Cortical GABA interneurons in neurovascular coupling: relays for subcortical vasoactive pathways. *J. Neurosci.* **24**, 8940–8949 (2004).
- Metea, M. R. & Newman, E. A. Glial cells dilate and constrict blood vessels: a mechanism of neurovascular coupling. *J. Neurosci.* **26**, 2862–2870 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Auger for initiating the experiments on cerebellar capillaries. This work was supported by the Wellcome Trust, the EU and a Wolfson-Royal Society Award. C.H. is in the 4-year PhD Programme in Neuroscience at UCL.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.A. (D.Attwell@ucl.ac.uk).

Fast neurotransmitter release triggered by Ca influx through AMPA-type glutamate receptors

Andrés E. Chávez¹, Joshua H. Singer^{1†} & Jeffrey S. Diamond¹

Feedback inhibition at reciprocal synapses between A17 amacrine cells and rod bipolar cells (RBCs) shapes light-evoked responses in the retina^{1–3}. Glutamate-mediated excitation of A17 cells elicits GABA (γ -aminobutyric acid)-mediated inhibitory feedback onto RBCs^{4–6}, but the mechanisms that underlie GABA release from the dendrites of A17 cells are unknown. If, as observed at all other synapses studied, voltage-gated calcium channels (VGCCs) couple membrane depolarization to neurotransmitter release⁷, feed-forward excitatory postsynaptic potentials could spread through A17 dendrites to elicit ‘surround’ feedback inhibitory transmission at neighbouring synapses. Here we show, however, that GABA release from A17 cells in the rat retina does not depend on VGCCs or membrane depolarization. Instead, calcium-permeable AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptors (AMPA), activated by glutamate released from RBCs, provide the calcium influx necessary to trigger GABA release from A17 cells. The AMPAR-mediated calcium signal is amplified by calcium-induced calcium release (CICR) from intracellular calcium stores. These results describe a fast synapse that operates independently of VGCCs and membrane depolarization and reveal a previously unknown form of feedback inhibition within a neural circuit.

In the mammalian retina, RBC synaptic terminals receive inhibitory synapses from several types of amacrine cell^{8,9} including GABAergic A17 amacrine cells^{5,9–12}, which receive excitatory input from RBCs and make immediately adjacent synapses back onto the same terminals. To investigate how A17 cells transduce RBC input into reciprocal release of GABA, we recorded synaptic responses from both cell types in acute retinal slices, beginning with RBCs (Fig. 1a). Depolarization of voltage-clamped RBCs elicited sustained inward calcium currents, upon which were superimposed transient, feedback inhibitory postsynaptic currents (IPSCs)^{5,6,13,14} that were reduced only slightly by the GABA_C receptor (GABA_CR) antagonist (1,2,5,6-tetrahydropyridin-4-yl) methylphosphonic acid (TPMPA, 50 μ M; $P = 0.012$), but were blocked by the GABA_A receptor (GABA_AR) antagonist SR95531 (10 μ M; $P = 5.3 \times 10^{-5}$; Fig. 1b, d). These voltage-step-evoked IPSCs (vIPSCs) were nearly completely resistant to the sodium channel blocker tetrodotoxin (TTX, 0.5 μ M; $P = 0.036$; Fig. 1c, d), which eliminates feedback inhibition from many amacrine cells^{15,16}. TTX has been reported to exert variable effects on reciprocal feedback in RBCs^{5,14}, but it does not affect voltage-gated conductances in rat RBCs and A17 cells^{5,17}. The vIPSCs were abolished by 5,7-dihydroxytryptamine (DHT, 50 μ M; $P = 0.0012$; Fig. 1c, d), which kills indoleamine-accumulating cells when taken up and oxidized by monoamine oxidase (MAO)¹⁸ and ablates A17 cells selectively when injected intravitreally several weeks before enucleation^{1,3}. Here, bath-applied DHT acted rapidly and specifically (see Supplementary Fig. 1): DHT eliminated excitatory postsynaptic currents (EPSCs) in A17 cells within 5 min but barely affected EPSCs in AII amacrine cells, which also receive synaptic

input from RBCs; DHT reduced calcium currents only slightly and did not affect the responses of RBCs to GABA. The acute effects of DHT on A17 EPSCs were blocked by the MAO inhibitor phenelzine (10 μ M; see Supplementary Fig. 1d–f).

IPSCs were also elicited in RBCs by stimulating amacrine cell dendrites directly with brief puffs of glutamate (50 μ M, 25 ms) delivered to the inner plexiform layer (IPL), where RBCs, A17 cells and other amacrine cells make synaptic contact (Fig. 1e, f). Glutamate-evoked IPSCs (gIPSCs) reversed near to the calculated chloride equilibrium potential ($E_{Cl} = -40$ mV; Fig. 1f inset) and contained components mediated by GABA_CRs and GABA_ARs that were blocked by TPMPA (50 μ M) and SR95531 (10 μ M), respectively⁴ (Fig. 1f, h). Glutamate puffs in the outer plexiform layer (OPL) elicited no responses in RBCs (see Supplementary Fig. 2), indicating that gIPSCs reflected GABA release in the IPL. TTX partially blocked gIPSCs ($P = 0.0008$; Fig. 1g, h), confirming that TTX-sensitive amacrine cells also provide (non-reciprocal) inhibition to RBCs^{15,16}. The TTX-insensitive component of the gIPSC required GABA release from A17 cells, as it was abolished by DHT ($P = 4 \times 10^{-5}$; Fig. 1g, h). DHT applied alone blocked only a fraction of the gIPSC ($P = 0.002$; Fig. 1h), indicating that it did not eliminate transmission from GABAergic amacrine cells other than A17 cells. The TTX-insensitive component of the gIPSC was partially sensitive to TPMPA ($56 \pm 6\%$ of control, $n = 4$, $P = 0.0015$), consistent with previous evidence that activation of A17 cells by multiple RBCs³—or enhanced activation of A17 cells by a single RBC^{5,6}—recruits a GABA_CR component in the feedback IPSC. To isolate the A17-mediated component of the gIPSC, all subsequent experiments were performed in the presence of TTX, except where noted.

At reciprocal synapses in the olfactory bulb, NMDA (*N*-methyl-D-aspartate) receptors (NMDARs) activated by glutamate released from mitral cells help to trigger GABAergic feedback¹⁹. In contrast, we found that feedback from A17 cells to RBCs requires activation of AMPARs but not NMDARs. Feedback IPSCs were blocked by philanthotoxin 433 (PhTx, 1 μ M), a specific antagonist of calcium-permeable AMPARs²⁰ (vIPSC: $P = 5 \times 10^{-5}$; gIPSC: $P = 0.0003$; Fig. 2a, b, g). PhTx did not act by blocking nicotinic acetylcholine receptors (nAChRs), because the nAChR antagonist d-tubocurarine (dTC, 20 μ M) did not reduce vIPSCs ($P = 0.16$; Fig. 2g). vIPSCs also were eliminated by the AMPAR/kainate receptor antagonist 2,3-dihydroxy-6-nitro-7-sulphamoyl-benzo(f)quinoxaline (NBQX, 50 μ M; vIPSC: $P = 2.4 \times 10^{-4}$; gIPSC: $P = 0.002$; Fig. 2g) and the AMPAR antagonist GYKI 53655 (25 μ M; vIPSC: $P = 8.1 \times 10^{-5}$; gIPSC: $P = 4.7 \times 10^{-4}$; Fig. 2g), but the NMDAR antagonist 3-(2-carboxypiperazin-4-yl)propyl-1-phosphonic acid (CPP, 10 μ M) did not reduce feedback IPSCs (vIPSC: $P = 0.051$; gIPSC: $P = 0.17$; Fig. 2c, d, g). EPSCs recorded in A17 cells, evoked by depolarizing a population of ON bipolar cells (including RBCs) with puff application of the group II/III mGluR antagonist (RS)- α -cyclopropyl-4-phosphonophenylglycine (CPPG, 600 μ M, 100 ms)²¹ in the OPL,

¹Synaptic Physiology Unit, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892-3701, USA. [†]Present address: Northwestern University, Departments of Ophthalmology and Physiology, 303 E. Chicago Avenue, Tarry Building, 5-727, Chicago, Illinois 60611, USA.

were blocked by PhTx ($P = 0.009$; Fig. 2e, g) and showed inward rectification (Fig. 2f), indicating that they were mediated by calcium-permeable AMPARs. These results corroborate previous physiological data indicating that AMPARs, not NMDARs, mediate synaptic inputs to A17 cells in rat retina^{5,6,17}, in contrast to teleost retina, where NMDARs contribute to reciprocal feedback onto RBCs¹⁴.

Calcium-permeable AMPARs could contribute to reciprocal GABA release by depolarizing A17 dendrites to activate VGCCs and/or by providing calcium directly. To distinguish between these possibilities, recordings were made from synaptically coupled RBC–A17 cell pairs (Fig. 3A, B). Depolarizing the RBC elicited an EPSC in the A17 cell and quantal feedback IPSCs in the RBC (Fig. 3A), but depolarizing the A17 cell evoked no detectable IPSC in the RBC ($n = 7$; Fig. 3B), indicating that membrane depolarization triggers release from RBCs but not A17 cells. Moreover, the

VGCC blocker cadmium (Cd, 200 μM) reduced the VGCC current in RBCs and abolished vIPSCs ($P = 9 \times 10^{-6}$; Fig. 3C, I) but did not affect gIPSCs ($P = 0.08$; Fig. 3D, I). Similar effects on gIPSCs were obtained with nimodipine (10 μM ; $P = 0.09$; Fig. 3I) or kurtaxin (350 nM; Fig. 3I), which together inhibit voltage-activated conformational changes in most subtypes of calcium channel^{22,23}. Nimodipine alone blocked calcium currents evoked by steps from -60 to -10 mV in RBCs ($P = 0.0013$; Fig. 3I). GABA release from A17 cells does require calcium influx, because gIPSCs were reduced when extracellular calcium was lowered from 2.5 mM to 0.5 mM ($P = 0.0013$; Fig. 3E, I) and were abolished when calcium was removed completely (0 mM Ca, 1.5 mM EGTA; $P = 2.4 \times 10^{-5}$; Fig. 3I). gIPSCs were also reduced after bath application of the membrane-permeant calcium chelator BAPTA-AM (50 μM ; $P = 0.0004$; Fig. 3F, I). These results indicate that AMPARs, not VGCCs, mediate the calcium influx that is required for GABA release from A17 cells. In contrast, glycinergic gIPSCs in RBCs were abolished by Cd (see Supplementary Fig. 3), indicating that other amacrine cells use VGCCs to trigger transmitter release.

gIPSCs could appear insensitive to VGCC blockade if exogenous glutamate activated AMPARs so strongly that consequent calcium influx saturated the release machinery, rendering VGCCs unnecessary. However, Cd did not reduce gIPSCs even in the presence of low extracellular calcium ($94 \pm 5\%$ of low calcium alone, $n = 5$, $P = 0.1$; Fig. 3E), indicating that VGCCs do not contribute even when the

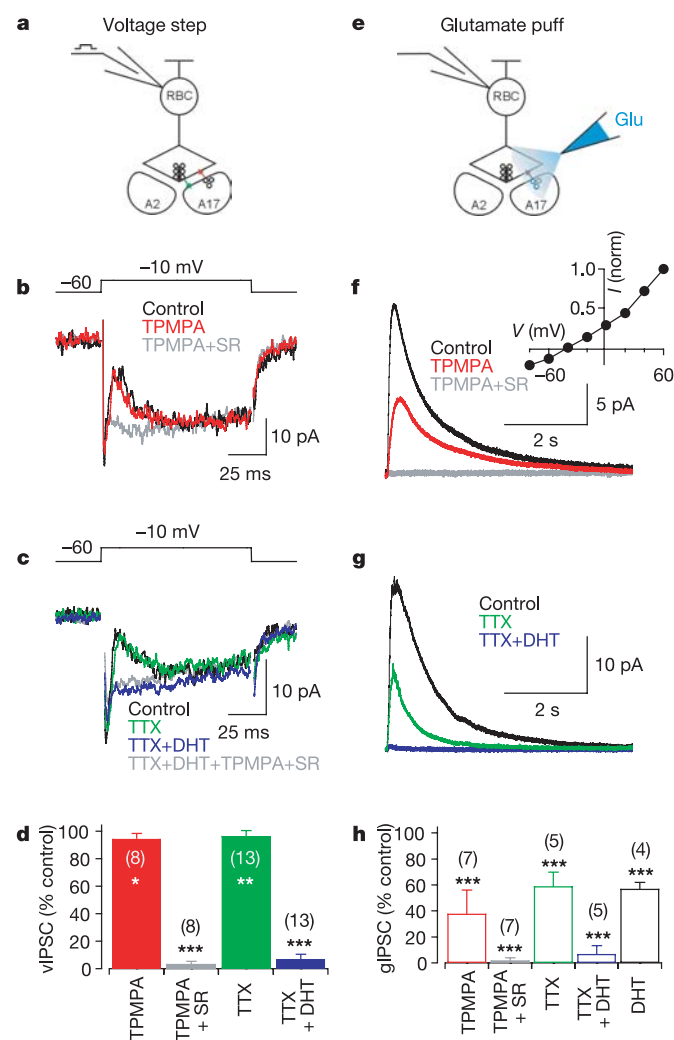


Figure 1 | Reciprocal GABA-mediated synaptic input to RBCs is mediated by A17 amacrine cells. **a**, Diagram of experiments in which feedback inhibition was elicited by depolarizing the RBC. A2 and A17 are amacrine cells. **b**, vIPSCs are only slightly reduced by the GABA_CR antagonist TPMPA (50 μM) and blocked by the GABA_AR antagonist SR95531 (10 μM). **c**, vIPSCs are diminished slightly by TTX (0.5 μM) and abolished by DHT (50 μM), which selectively ablates A17 cells. GABA_AR antagonists exerted no further effect after DHT application. **d**, Summarized pharmacological profile of vIPSCs. **e**, Diagram of experiments in which A17 cells were activated by exogenous application of glutamate. **f**, gIPSCs are sensitive to both TPMPA (50 μM) and SR95531 (10 μM). Inset, current–voltage relationship of gIPSCs. **g**, gIPSCs are partially sensitive to TTX (0.5 μM), and the remaining component is blocked by DHT (50 μM). **h**, Summarized pharmacological profile of gIPSCs.

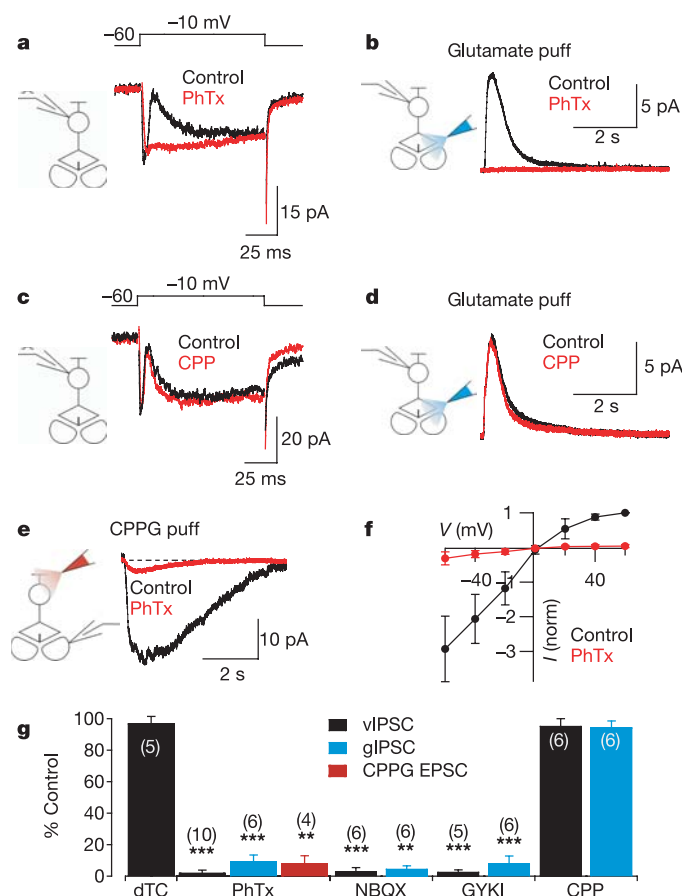


Figure 2 | Calcium-permeable AMPARs trigger GABA release from A17 cells. **a**, vIPSCs are blocked by the calcium-permeable AMPAR antagonist PhTx (1 μM). **b**, gIPSCs are also blocked by PhTx. **c**, vIPSCs are unaffected by the NMDAR antagonist CPP (10 μM). **d**, gIPSCs are also unaffected by CPP. **e**, EPSCs recorded in A17 cells ($V_{\text{hold}} = -60$ mV) and elicited by depolarizing RBCs with the mGluR antagonist CPPG are reduced by PhTx (1 μM). **f**, Current–voltage plot indicating voltage dependence and PhTx-sensitivity of EPSCs in A17 cells. EPSC amplitudes were normalized to control response at $+60$ mV. **g**, Summarized drug effects.

GABA release machinery is not activated maximally. Moreover, IPSCs elicited by puffs of artificial cerebrospinal fluid (ACSF) containing 110 mM potassium (kIPSCs), which would depolarize both RBC and A17 processes, were abolished by NBQX ($P = 0.002$; Fig. 3G, I). If GABA release were evoked by membrane depolarization, kIPSCs would have been partially resistant to AMPAR blockade. Finally, direct depolarization of voltage-clamped A17 cells failed to elicit IPSCs in synaptically coupled RBCs (Fig. 3B). EPSCs recorded in A17 cells were reversed at positive holding potentials (Fig. 2f), indicating that a patch electrode on an A17 cell soma could depolarize the postsynaptic membrane potential sufficiently to

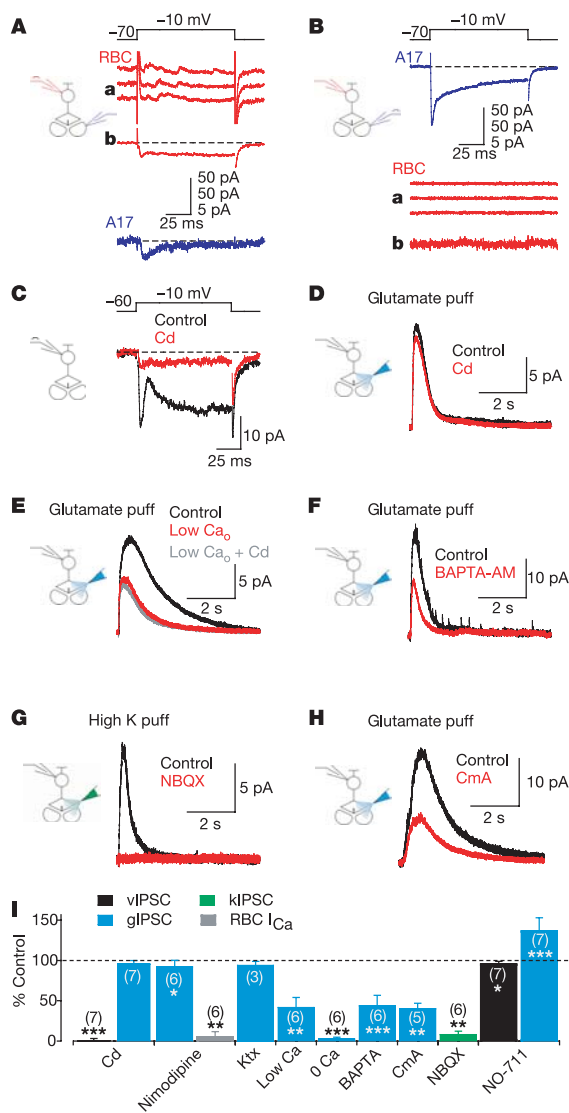


Figure 3 | Neither membrane depolarization nor VGCC activation triggers GABA release from A17 cells. **A**, RBC depolarization elicits an EPSC in a coupled A17 cell ($V_{hold} = -70$ mV). Traces in **a** show single RBC responses; trace in **b** shows a leak-subtracted ($P/4$) average RBC response. **B**, Depolarizing the same A17 cell elicits no response in the same RBC ($V_{hold} = 0$ mV; 0.2 mM EGTA in A17). Traces in **a** show single RBC responses; trace in **b** shows average RBC response. **C**, Cd (200 μM) reduces VGCC-mediated current in RBCs and eliminates vIPSCs. **D**, Cd does not affect gIPSCs. **E**, gIPSCs are reduced when extracellular calcium is lowered from 2.5 mM (control) to 0.5 mM. Cd does not reduce GABA release even in low extracellular calcium. **F**, gIPSCs are reduced by the calcium chelator BAPTA-AM (50 μM). **G**, Puff application of high-potassium (110 mM) ACSF elicits a feedback IPSC (kIPSC) that is blocked completely by NBQX, indicating that direct depolarization of A17 cells does not elicit GABA release. **H**, gIPSCs are reduced when vesicle filling is inhibited with CmA. **I**, Summarized drug effects.

activate any VGCCs, if they were present at A17 GABA release sites.

Previous evidence indicates that GABA release from A17 cells is vesicular: Postsynaptic processes on A17 cells contain small clear synaptic vesicles^{9,24}, indoleamine-accumulating (A17-like) postsynaptic processes in the rabbit retina contain synaptic vesicle proteins²⁵, and vIPSCs in RBCs exhibit quantal components^{4–6} (Fig. 3A). In addition, gIPSCs were reduced significantly when vacuolar proton-translocating ATPases (and, therefore, vesicular filling) were inhibited by bath application of concanamycin A (CmA, 10 μM; $P = 0.004$; Fig. 3H, I)²⁶. Reversed uptake does not contribute substantially to GABA release, as the neuronal GABA transporter blocker NO-711 (10 μM) reduced vIPSCs only slightly ($P = 0.02$; Fig. 3I) and enhanced gIPSCs ($P = 0.0008$; Fig. 3I). This latter effect probably reflected slowed removal of synaptically released GABA, as NO-711 enhanced RBC responses to exogenous GABA to a similar extent ($135 \pm 21\%$, $n = 5$, $P = 0.02$; data not shown).

CICR from intracellular stores in A17 cells could enhance coupling between AMPAR activation and GABA release by amplifying the postsynaptic calcium signal^{27,28}. Accordingly, activating the release of calcium from intracellular stores by puff application of the ryanodine receptor (RyR) agonist caffeine²⁷ (15 mM, 50 ms) in the IPL elicited IPSCs in RBCs that were blocked by GABA antagonists ($P = 0.002$; Fig. 4g), DHT ($P = 1.2 \times 10^{-5}$; Fig. 4a, g), or the RyR antagonist

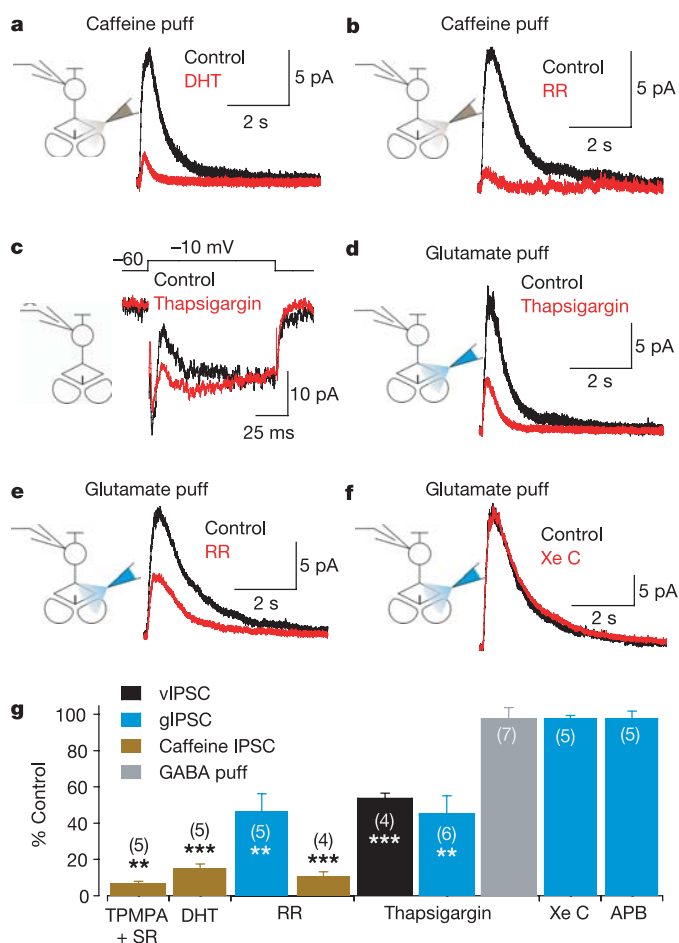


Figure 4 | Calcium signalling in A17 amacrine cells is amplified by CICR. **a**, Activation of RyRs by caffeine (15 mM) elicits IPSCs in RBCs that are blocked by DHT. **b**, Caffeine-evoked IPSCs are blocked by the RyR antagonist RR (40 μM). **c**, vIPSCs are reduced when intracellular calcium stores are reduced by bath application of thapsigargin (1 μM; thapsigargin was also included in the patch pipette). **d**, Thapsigargin reduces the gIPSC. **e**, RR partially reduces gIPSCs. **f**, Xestospingon C (Xe C, 3 μM), an Ins(1,4,5)P₃ receptor antagonist, does not affect glutamate-evoked IPSCs. **g**, Summarized drug effects.

ruthenium red²⁷ (RR, 40 μ M; $P = 0.0005$; Fig. 4b, g). In addition, vIPSCs and gIPSCs were diminished when intracellular calcium stores were reduced by bath application of 1 μ M thapsigargin (vIPSC: $P = 0.004$; gIPSC: $P = 0.006$; Fig. 4c, d, g). Thapsigargin did not affect the responses of RBCs to exogenously applied GABA (100 μ M, 25 ms; $P = 0.59$; Fig. 4g). RR blocked the gIPSC only partially ($P = 0.002$; Fig. 4e, g), indicating that some of the calcium that contributes to GABA release might arise from another source, perhaps inositol-1,4,5-trisphosphate (Ins(1,4,5)P₃)-receptor-sensitive stores²⁸ or influx through calcium-permeable AMPARs. gIPSCs were unaffected by the Ins(1,4,5)P₃ receptor antagonists xestospongins C (3 μ M; $P = 0.06$; Fig. 4f, g) or 2-APB (50 μ M; $P = 0.3$; Fig. 4g), arguing against a role for Ins(1,4,5)P₃ receptor-operated stores. Thus, AMPAR-mediated calcium influx into A17 cells triggers GABA release directly and through CICR.

Our results reveal a fast synapse at which AMPARs provide calcium influx to trigger neurotransmitter release, and at which release occurs independently of membrane depolarization. At most synapses, VGCCs couple presynaptic membrane potential and neurotransmitter release⁷. In many cells, postsynaptic depolarization recruits VGCCs to mediate associative interactions between the soma and the dendritic arborization²⁹ or, in the olfactory bulb, to trigger reciprocal GABA release¹⁹. Although specific physiological roles for VGCCs in A17 cells remain unclear, they might replenish intracellular calcium stores or trigger the release of other neurotransmitters.

If GABA release from A17 cells were triggered by VGCCs, the spread of depolarization through the A17 dendrites could contribute to surround inhibition by eliciting release at electrotonically adjacent synapses²⁵. By using calcium-permeable AMPARs rather than VGCCs to trigger GABA release, however, A17 dendrites might compartmentalize reciprocal feedback to maintain synapse specificity. The rapid amplification of the signal by CICR probably occurs in the immediate vicinity of the postsynaptic membrane and is unlikely to compromise this specificity, particularly if, as in the rabbit²⁵, reciprocal synapses in rat A17 dendrites are separated by 20 μ m or more. Further experiments are required to determine the spatial extent of synaptic calcium signalling in A17 dendrites.

At other reciprocal synapses in the olfactory bulb and goldfish retina, NMDAR activation triggers GABAergic feedback that is much slower than that observed here^{14,19}. Feedback in A17 cells might require calcium-permeable AMPARs, which show faster kinetics than NMDARs, to confer transience on the visual signal^{1–3} and to prevent rapid depletion of the readily-releasable vesicle pool in RBC terminals³⁰.

METHODS

See Supplementary Information for a detailed description of methods. Retinal slices were prepared from Sprague–Dawley rats (P17–21) as described previously⁶. Except where indicated, experiments were performed in solutions supplemented with strychnine (1 μ M) and tetrodotoxin (TTX, 0.5 μ M) to block glycine receptors and voltage-gated sodium channels, respectively.

During whole-cell recordings, RBCs and A17 cells were filled with Alexa-488 through the patch pipette and visualized by epifluorescent illumination to confirm cell type. vIPSC amplitude was measured by fitting the last 30–50 ms of the current during the voltage response to a straight line, extrapolating the line to the time point of the IPSC peak and measuring the difference (see Supplementary Fig. 4). Other responses were measured as the difference between the peak and the baseline before the stimulus. Unless indicated otherwise, statistical comparisons were made with a paired, two-tailed Student's *t*-test (Igor Pro) and significance was concluded when $P < 0.05$. In the figures, * indicates $P < 0.05$, ** indicates $P < 0.01$ and *** indicates $P < 0.001$. Data are reported as mean $\pm$ s.d.

Received 10 May; accepted 31 July 2006.

Published online 1 October 2006.

1. Nakatsuka, K. & Hamasaki, D. I. Destruction of the indoleamine-accumulating amacrine cells alters the ERG of rabbits. *Invest. Ophthalmol. Vis. Sci.* **26**, 1109–1116 (1985).
2. Euler, T. & Masland, R. H. Light-evoked responses of bipolar cells in a mammalian retina. *J. Neurophysiol.* **83**, 1817–1829 (2000).

3. Dong, C. J. & Hare, W. A. Temporal modulation of scotopic visual signals by A17 amacrine cells in mammalian retina *in vivo*. *J. Neurophysiol.* **89**, 2159–2166 (2003).
4. Hartveit, E. Membrane currents evoked by ionotropic glutamate receptor agonists in rod bipolar cells in the rat retinal slice preparation. *J. Neurophysiol.* **76**, 401–422 (1996).
5. Hartveit, E. Reciprocal synaptic interactions between rod bipolar cells and amacrine cells in the rat retina. *J. Neurophysiol.* **81**, 2923–2936 (1999).
6. Singer, J. H. & Diamond, J. S. Sustained Ca²⁺ entry elicits transient postsynaptic currents at a retinal ribbon synapse. *J. Neurosci.* **23**, 10923–10933 (2003).
7. Katz, B. & Miledi, R. The timing of calcium action during neuromuscular transmission. *J. Physiol. (Lond.)* **189**, 535–544 (1967).
8. Kolb, H. & Nelson, R. Amacrine cells of the cat retina. *Vision Res.* **21**, 1625–1633 (1981).
9. Sterling, P. & Lampson, L. A. Molecular specificity of defined types of amacrine synapse in cat retina. *J. Neurosci.* **6**, 1314–1324 (1986).
10. Nelson, R. & Kolb, H. A17: a broad-field amacrine cell in the rod system of the cat retina. *J. Neurophysiol.* **54**, 592–614 (1985).
11. Sandell, J. H. & Masland, R. H. A system of indoleamine-accumulating neurons in the rabbit retina. *J. Neurosci.* **6**, 3331–3347 (1986).
12. Vaney, D. I. Morphological identification of serotonin-accumulating neurons in the living retina. *Science* **233**, 444–446 (1986).
13. Singer, J. H., Lassova, L., Vardi, N. & Diamond, J. S. Coordinated multivesicular release at a mammalian ribbon synapse. *Nature Neurosci.* **7**, 826–833 (2004).
14. Vigh, J. & von Gersdorff, H. Prolonged reciprocal signalling via NMDA and GABA receptors at a retinal ribbon synapse. *J. Neurosci.* **25**, 11412–11423 (2005).
15. Bloomfield, S. A. & Xin, D. Surround inhibition of mammalian AII amacrine cells is generated in the proximal retina. *J. Physiol. (Lond.)* **523**, 771–783 (2000).
16. Shields, C. R. & Lukasiewicz, P. D. Spike-dependent GABA inputs to bipolar cell axon terminals contribute to lateral inhibition of retinal ganglion cells. *J. Neurophysiol.* **89**, 2449–2458 (2003).
17. Menger, N. & Wässle, H. Morphological and physiological properties of the A17 amacrine cell of the rat retina. *Vis. Neurosci.* **17**, 769–780 (2000).
18. Baumgarten, H. G. et al. Mode and mechanism of action of neurotoxic indoleamines: a review and a progress report. *Ann. NY Acad. Sci.* **305**, 3–24 (1978).
19. Schoppa, N. E. & Urban, N. N. Dendritic processing within olfactory bulb circuits. *Trends Neurosci.* **26**, 501–506 (2003).
20. Washburn, M. S. & Dingledine, R. Block of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors by polyamines and polyamine toxins. *J. Pharmacol. Exp. Ther.* **278**, 669–678 (1996).
21. Nawy, S. Desensitization of the mGluR6 transduction current in tiger salamander On bipolar cells. *J. Physiol. (Lond.)* **558**, 137–146 (2004).
22. Hess, P., Lansman, J. B. & Tsien, R. W. Different modes of Ca channel gating behaviour favoured by dihydropyridine Ca agonists and antagonists. *Nature* **311**, 538–544 (1984).
23. Sidach, S. S. & Mintz, I. M. Kurtosis, a gating modifier of neuronal high- and low-threshold Ca channels. *J. Neurosci.* **22**, 2023–2034 (2002).
24. Brandstätter, J. H., Koulen, P. & Wässle, H. Selective synaptic distribution of kainate receptor subunits in the two plexiform layers of the rat retina. *J. Neurosci.* **17**, 9298–9307 (1997).
25. Zhang, J., Li, W., Trexler, E. B. & Massey, S. C. Confocal analysis of reciprocal feedback at rod bipolar terminals in the rabbit retina. *J. Neurosci.* **22**, 10871–10882 (2002).
26. Drose, S. & Altendorf, K. Bafilomycins and concanamycins as inhibitors of V-ATPases and P-ATPases. *J. Exp. Biol.* **200**, 1–8 (1997).
27. Verkhratsky, A. & Petersen, O. H. The endoplasmic reticulum as an integrating signalling organelle: from neuronal signalling to neuronal death. *Eur. J. Pharmacol.* **447**, 141–154 (2002).
28. Warrior, A., Borges, S., Dalcino, D., Walters, C. & Wilson, M. Calcium from internal stores triggers GABA release from retinal amacrine cells. *J. Neurophysiol.* **94**, 4196–4208 (2005).
29. Sjostrom, P. J. & Nelson, S. B. Spike timing, calcium signals and synaptic plasticity. *Curr. Opin. Neurobiol.* **12**, 305–314 (2002).
30. Singer, J. H. & Diamond, J. S. Vesicle depletion and synaptic depression at a mammalian ribbon synapse. *J. Neurophysiol.* **95**, 3191–3198 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Swartz for his gift of kurtosis, J. Isaac for his gift of GYKI 53655, and J. Isaac, D. Copenhagen, C. Jahr and members of the Diamond laboratory for comments on the manuscript. This research was supported by the NINDS Intramural Research Program and a K22 award to J.H.S. A.E.C. is a doctoral student in a graduate program partnership between NIH and the University of Valparaíso, Chile.

Author Contributions A.E.C. and J.H.S. collected and analysed data and helped to design experiments; J.S.D. directed the study, helped to design experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.S.D. (diamondj@ninds.nih.gov).

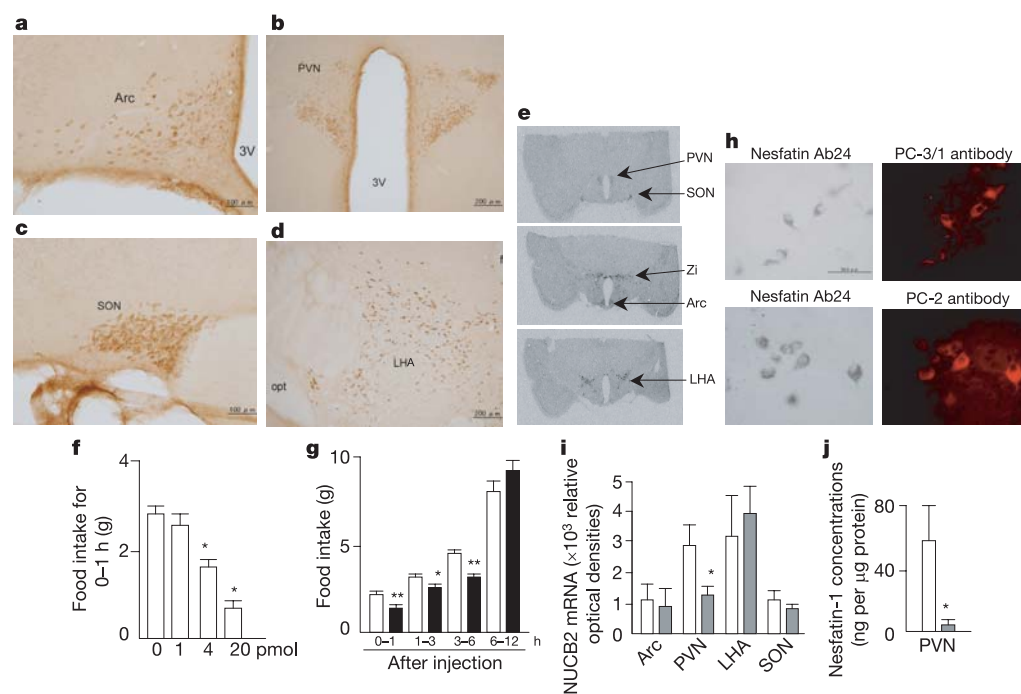
Identification of nesfatin-1 as a satiety molecule in the hypothalamus

Shinsuke Oh-I¹, Hiroyuki Shimizu¹, Tetsurou Satoh¹, Shuichi Okada¹, Sachika Adachi², Kinji Inoue², Hiroshi Eguchi³, Masanori Yamamoto³, Toshihiro Imaki¹, Koushi Hashimoto¹, Takafumi Tsuchiya¹, Tsuyoshi Monden¹, Kazuhiko Horiguchi¹, Masanobu Yamada¹ & Masatomo Mori¹

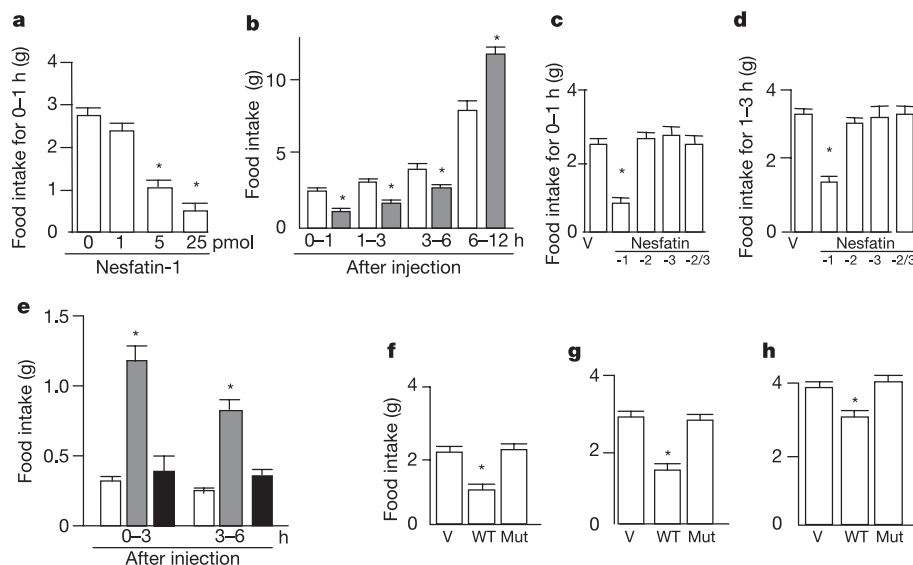
The brain hypothalamus contains certain secreted molecules that are important in regulating feeding behaviour^{1–3}. Here we show that nesfatin, corresponding to NEFA/nucleobindin2 (NUCB2), a secreted protein of unknown function, is expressed in the appetite-control hypothalamic nuclei in rats. Intracerebroventricular (i.c.v.) injection of NUCB2 reduces feeding. Rat cerebrospinal fluid contains nesfatin-1, an amino-terminal fragment derived from NUCB2, and its expression is decreased in the hypothalamic paraventricular nucleus under starved conditions. I.c.v. injection of nesfatin-1 decreases food intake in a dose-dependent manner, whereas injection of an antibody neutralizing nesfatin-1 stimulates appetite. In contrast, i.c.v. injection of other possible fragments processed from NUCB2 does not promote satiety, and conversion of NUCB2 to nesfatin-1 is necessary to induce feeding suppression. Chronic i.c.v. injection of nesfatin-1 reduces body weight, whereas rats gain body weight after chronic i.c.v. injection of antisense morpholino oligonucleotide against the gene encoding NUCB2. Nesfatin-1-induced anorexia occurs in Zucker rats with a leptin receptor mutation, and an anti-nesfatin-1 antibody

does not block leptin-induced anorexia. In contrast, central injection of α -melanocyte-stimulating hormone elevates NUCB2 gene expression in the paraventricular nucleus, and satiety by nesfatin-1 is abolished by an antagonist of the melanocortin-3/4 receptor. We identify nesfatin-1 as a satiety molecule that is associated with melanocortin signalling in the hypothalamus.

We attempted to identify any new appetite-regulating molecules by using a subtraction-cloning assay of peroxisome proliferator-activated receptor- γ activator (troglitazone)-stimulated genes in SQ-5 cells. First, we used a computer-based program (Signal Peptide Prediction, http://bioinformatics.leeds.ac.uk/prot_analysis/Signal.html) to search for a signal peptide in the coding region of each of the 596 genes cloned in this assay. As most hypothalamic secreted molecules also exist in peripheral adipose tissue⁴, we examined whether the genes cloned were expressed in both brain medulloblastoma (HTB185) and 3T3-L1 adipocyte cells. Nine genes meeting these criteria were identified, and the expression of one gene among those identified was profoundly stimulated by troglitazone in SQ-5 cells (Supplementary Fig. 1). Sequencing demonstrated that this



¹Department of Medicine and Molecular Science, Gunma University Graduate School of Medicine, Showa-machi, Maebashi 371-8511, Japan. ²Department of Regulation Biology, Saitama University, Shimo-okubo, Saitama 378-8570, Japan. ³Pharmaceutical Discovery Research Laboratory, Teijin Pharma Limited, Asahigaoka, Hino 191-8512, Japan.

**Figure 2 | Nesfatin-1-induced satiety in rats.**

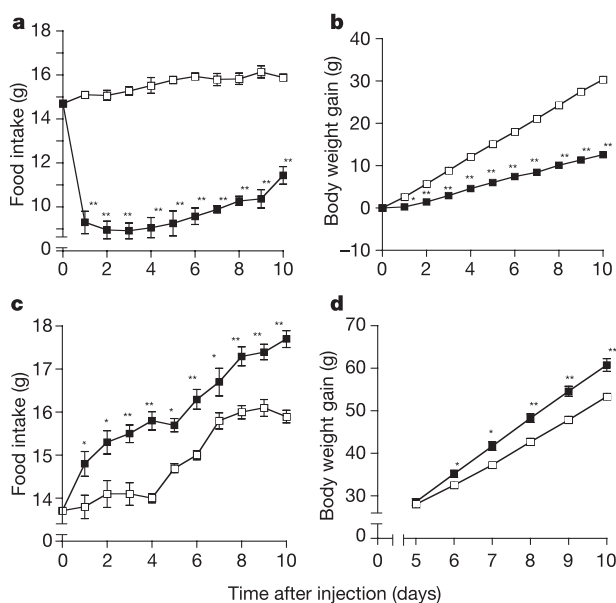
a, b, Food intake ($n = 5-8$) after i.c.v. injection of nesfatin-1; **a**, dose-response; **b**, time course (open bars, 0 pmol nesfatin-1; filled bars, 5 pmol nesfatin-1). **c, d**, Food intake ($n = 5-10$) after i.c.v. injection of each fragment (25 pmol) derived from NUCB2: **c**, 0-1 h; **d**, 1-3 h. **e**, Food intake ($n = 5-7$) after i.c.v. injection of 8 µg IgG (open bars, control rabbit IgG; grey bars, nesfatin Ab24 IgG; black bars, nesfatin Ab301 IgG). **f-h**, Food intake ($n = 5-6$) after i.c.v. injection of each intermediate (5 pmol) with wild-type Lys 83-Arg 84 (WT) or mutant Ala 83-Ala 84 (Mut). **f**, 0-1 h; **g**, 1-3 h; **h**, 3-6 h. Data are means $\pm$ s.e.m. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ compared with 0 pmol, vehicle (V) or control IgG.

complementary DNA fragment corresponds to the 5'-untranslated region of the gene encoding DNA binding/EF-hand/acidic protein (NEFA) or NUCB2. NUCB2 is composed of a signal peptide of 24 amino acids and a protein structure containing 396 amino acids⁵. The homology of the amino-acid sequence of NUCB2 is highly conserved in humans, mice and rats (87.4% homology to humans and 95.7% to mice). Because NUCB2 belongs to a homologous gene family with nucleobindin1 (NUCB1)⁶, these two genes might have arisen from a single EF-hand ancestor⁷. NUCB2 and NUCB1 are secreted proteins⁵⁻⁸ but their function remains unknown, although these proteins have been suggested to contribute to bone mineralization and antibody production^{9,10}.

Troglitazone does not cross the blood-brain barrier to alter feeding behaviour¹¹ and does not change body weights in rodents (Supplementary Fig. 2), but the present findings encouraged further investigation into whether the protein deduced from this identified gene is present in the hypothalamus. To address this issue, an anti-NUCB2 antibody (termed NUCB2 Ab-L) was used to perform immunohistochemical analyses in rats. This antibody recognizes the whole structure of NUCB2, and its detection was not absorbed by the addition of appetite-regulating peptides (Supplementary Fig. 3). NUCB2 was expressed in the hypothalamic areas (Fig. 1a-d), namely the arcuate nucleus, paraventricular nucleus (PVN) and supraoptic nucleus, the lateral hypothalamic area and the zona incerta, which regulate feeding, and also in the nucleus tract solitarius (data not shown). Preincubation of the cognate peptide with NUCB2 Ab-L abolished staining (data not shown). Analysis of *in situ* hybridization demonstrated that the gene encoding NUCB2 is expressed in the same hypothalamic nuclei as those shown by the immunohistochemistry (Fig. 1e). We next identified that i.c.v. injection of recombinant protein of NUCB2 decreased food intake (Fig. 1f, g), but no apparent behavioural changes including increased locomotion or narcolepsy were noted during the experiments. Conversely, i.c.v. injection of IgG from NUCB2 Ab-L stimulated feeding, compared with that induced by control IgG (Supplementary Fig. 4). These results indicate that this protein is a hypothalamic endogenous molecule that induces anorexia. We therefore refer to this protein as nesfatin (for NEFA/nucleobindin2-encoded satiety- and fat-influencing protein).

Post-translational processing of a prohormone produces several fragments or peptides through cleavage at specific sites by prohormone convertases (PCs)¹², and rat NUCB2/nesfatin possesses potential cleavage sites that are flanked by a pair of basic amino acids, Lys-Arg or Arg-Arg (Supplementary Fig. 5). The ProP 1.0 prediction server¹³ predicted these sites and showed the highest score for cleavage at the

Lys 83-Arg 84 site, which is conserved among different species⁵⁻⁷. The possible processed fragments were designated as follows; nesfatin-1, residues 1-82; nesfatin-2, residues 85-163; and nesfatin-3, residues 166-396. The structures of nesfatin-2 and nesfatin-3 are characterized by their possessing putative DNA-binding and leucine-zipper sites that bind nucleosomal DNA^{5,6}, but nesfatin-1 does not preserve these sites. We produced two different types of antibody, nesfatin Ab24 and Ab301, which recognized synthetic nesfatin-1 and nesfatin-3, respectively (Supplementary Fig. 5). When monitored by nesfatin-1 ELISA, elution profiles of cerebrospinal fluid extracts on high-performance liquid

**Figure 3 | Body weight changes after continuous i.c.v. injection of substances.**

a, b, Daily food intake (**a**) and body weight gain (**b**; increment from day 0) in rats ($n = 4-5$) after nesfatin-1 injection (5 pmol daily). Open squares, vehicle; filled squares, nesfatin-1. **c, d**, Daily food intake (**c**) and body weight gain (**d**; increment from day 0) in rats ($n = 4-5$) after injection with NUCB2 missense (open squares) or antisense (filled squares) morpholino oligonucleotide (MON; 40 µg per day). Antisense (5'-ATGGTCTCCACCTCATCTTCAGAG-3') and 5-missense (5'-ATCGTGCTCCACGTCATCTACAGAG-3') MONs were purchased from Gene Tools. An Alzet osmotic mini-infusion pump was used for continuous injection. Data are means $\pm$ s.e.m. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ compared with vehicle or missense MON.

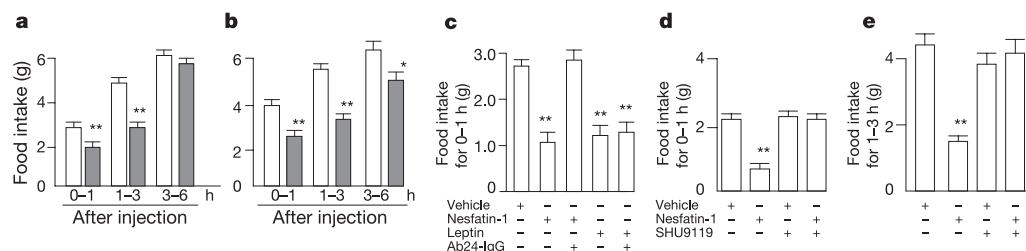


Figure 4 | Nesfatin-1-induced satiety associated with leptin or melanocortin signalling. **a, b**, Food intake in lean (**a**) and Zucker (**b**) rats ($n = 5$) after i.c.v. injection of 5 pmol nesfatin-1 (open bars, vehicle; filled bars, nesfatin-1). **c**, Effects of nesfatin Ab24 on leptin-induced anorexia ($n = 6$). Vehicle, 5 pmol nesfatin-1 or 5 pmol leptin was centrally

administered 15 min after i.c.v. injection of Ab24 IgG (8 μ g) during the dark phase. **d, e**, Effects of SHU9119 on nesfatin-1-induced anorexia ($n = 6$). Vehicle or 5 pmol nesfatin-1 was centrally administered 15 min after i.c.v. injection of 20 pmol SHU9119. Data are means $\pm$ s.e.m. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$ compared with vehicle.

chromatography showed an apparent peak corresponding to nesfatin-1 peptide (Supplementary Figs 6 and 7), indicating that nesfatin-1 is a secreted fragment. Immunohistochemical analysis with nesfatin Ab24 showed that nesfatin-1 was distributed in the hypothalamic nuclei in a similar manner to that observed for NUCB2 Ab-L (Supplementary Fig. 8) and was present in the cytoplasm, but not the nucleus, of the hypothalamic neurons, where it was co-localized with PC-3/1 and PC-2 (Fig. 1h). Under conditions of starvation for 24 h, NUCB2 gene expression was reduced in the PVN, in which nesfatin-1 concentration was also decreased (Fig. 1i, j). No significant changes were observed in other hypothalamic nuclei.

As a prohormone can be processed to generate bioactive fragments¹⁴, we next estimated the food intake of rats after i.c.v. injection of each of the possible fragments derived from NUCB2. Injection of a synthetic peptide identical to nesfatin-1 decreased food intake in a dose-dependent manner, and the effects continued for 6 h after injection (Fig. 2a, b). In contrast, neither a synthetic peptide corresponding to nesfatin-2, a fragment corresponding to nesfatin-3 nor a fragment corresponding to nesfatin-2/3 affected appetite (Fig. 2c, d). Subsequently, we found that food intake was significantly stimulated after injection of nesfatin Ab24-derived IgG, but not after injection of nesfatin Ab301-derived IgG (Fig. 2e). Finally, we found that although i.c.v. injection of the intermediate form (residues 1–223) of nesfatin retaining wild-type Lys 83–Arg 84 significantly reduced feeding, injection of the mutant form of nesfatin (Ala 83–Ala 84) did not induce anorexia (Fig. 2f–h), indicating that appetite suppression by NUCB2 requires conversion to nesfatin-1.

In the next experiments we estimated chronological changes in rat body weight. In comparison with rats receiving vehicle injection (Fig. 3a, b), rats receiving a continuous i.c.v. injection of nesfatin-1 unfailingly showed decreased food intake and suppressed body weight gain (at day 10, 30.4 ± 0.3 g versus 12.6 ± 0.7 g ($\pm$ s.e.m.), respectively). Nesfatin-1 injection also significantly decreased the weights of subcutaneous, epididymal and mesenteric fats, but not that of the gastrocnemius (Supplementary Fig. 9). Conversely, continuous i.c.v. injection of NUCB2 antisense morpholino oligonucleotide diminished the hypothalamic NUCB2 contents (Supplementary Fig. 10) and resulted in a consistent increase in appetite. Initially, body weight gains remained unchanged until 5 days after injection, but significant increases occurred after 6 days (Fig. 3c, d).

There are complex but integrated interconnections among the hypothalamic nuclei in regulating feeding^{1,2,15}. Leptin and pro-opiomelanocortin (POMC)-derived α -melanocyte-stimulating hormone are key anorectic molecules in the hypothalamus^{1–3}. In Zucker rats possessing a leptin receptor mutation¹⁶, significant nesfatin-1-induced satiety occurred (Fig. 4a, b). Previous injection of nesfatin Ab24 eliminated anorexia induced by nesfatin-1, but did not block leptin-induced anorexia (Fig. 4c), indicating a possible lack of involvement of hypothalamic leptin signalling in the induction of anorexia by nesfatin-1. In contrast, central injection of α -melanocyte-stimulating

hormone stimulated the expression of the PVN gene encoding NUCB2 (Supplementary Fig. 11), and previous injection of SHU9119, an antagonist specific for the melanocortin 3/4 receptor¹⁷, abolished nesfatin-1-induced satiety (Fig. 4d, e). However, injection of nesfatin-1 did not affect the expression of the genes encoding POMC, agouti-related peptide, neuropeptide Y or corticotropin-releasing hormone in the arcuate nucleus and PVN (Supplementary Fig. 12), and nesfatin-1 stimulated neither cAMP formation nor calcium influx in cells expressing the melanocortin-3 or -4 receptor (Supplementary Fig. 13). As the melanocortin-4 receptor is distributed widely in several brain regions that are plausibly involved in the coordinated control of feeding and energy expenditure^{15,18,19}, these findings imply the possible involvement of unknown neuronal substances, which are regulated by melanocortin-4 receptor signalling, in mediating nesfatin-1-induced anorexia. Taken together, the present studies indicate that hypothalamic nesfatin-1 signalling might involve the leptin-independent melanocortin signalling system²⁰.

We have identified a novel anorexigenic nesfatin corresponding to NUCB2 in the hypothalamus that produces a secreted fragment, nesfatin-1. At the carboxy terminus matching the nesfatin-2/3 fragment, NUCB2 possesses archetypal motifs, namely possible calcium-binding and DNA-binding sites^{5–7}. The C-terminal region also binds other molecules²¹. However, the present findings indicate that this C terminus is not involved in feeding regulation. In contrast, nesfatin-1 located in the N terminus of NUCB2 was indispensable to the induction of satiety, and conversion of NUCB2 to nesfatin-1 was essential to induce its activity *in vivo*.

The present findings indicate that nesfatin-1 might be a useful target for the development of drug therapies to treat obese persons.

METHODS

Further details are provided in Supplementary Information.

Subtraction-cloning assay. With the use of a polymerase chain reaction (PCR)-based subtraction cloning kit²², the troglitazone (100 μ M)-stimulated genes of SQ-5 cells derived from lung carcinomas, which express leptin and its receptor²³, were identified.

Proteins and substances. A cDNA encoding full-length mouse NUCB2 was subcloned into pMAL-c2x to yield a fusion protein with maltose binding protein in *Escherichia coli*. The intermediate forms of nesfatin retaining the wild-type motif Lys 83–Arg 84 or the mutant Lys 83–Arg 84 $\rightarrow$ Ala 83–Ala 84 (mutations generated by the PCR method) were constructed as glutathione S-transferase fusion proteins with a histidine tag in *E. coli* transfected with the pET41a(+) vector. Peptides for rat nesfatin-1 and nesfatin-2 were synthesized by Yanaihara Institute, Inc.

Antibodies. Antibodies were generated in rabbits against amino-acid sequences of rat NUCB2: Ab-L against residues 117–128, Ab24 against residues 24–38, and Ab301 against residues 301–314. A Cys residue was added to the C terminus of each peptide.

Animal experiments. Without anaesthesia, the test substance (5 μ l) was injected into the third ventricle of the brain in adult male Wistar rats weighing 200–250 g, as described previously²⁴. The experiments involving i.c.v. bolus injection were started 30 min before the beginning of the dark cycle (lighting from 6:00 to

18:00), and the experiments for IgG injection were started at 10:00 unless otherwise specified. Zucker obese rats were purchased from Charles River Laboratory.

Statistical analysis was performed by analysis of variance, unless otherwise mentioned.

Received 8 June; accepted 11 August 2006.

Published online 1 October 2006.

- Friedman, J. M. & Halaas, J. L. Leptin and the regulation of body weight in mammals. *Nature* **395**, 763–770 (1998).
- Schwartz, M. W. *et al.* Central nervous system control of food intake. *Nature* **404**, 661–671 (2000).
- Stanley, S., Wynne, K., McGowan, B. & Bloom, S. Hormonal regulation of food intake. *Physiol. Rev.* **85**, 1131–1158 (2005).
- Shimizu, H. & Mori, M. The brain–adipose axis: a review of involvement of molecules. *Nutr. Neurosci.* **8**, 7–20 (2005).
- Barnikol-Watanabe, S. *et al.* Human protein NEFA, a novel DNA binding/EF-hand/leucine zipper protein. Molecular cloning and sequence analysis of the cDNA, isolation and characterization of the protein. *Biol. Chem. Hoppe-Seyler* **375**, 497–512 (1994).
- Miura, K., Titani, K., Kurisawa, Y. & Kanai, Y. Molecular cloning of nucleobindin, a novel DNA-binding protein that contains both a signal peptide and a leucine zipper structure. *Biochem. Biophys. Res. Commun.* **187**, 375–380 (1992).
- Karabinos, A. *et al.* The divergent domains of the NEFA and nucleobindin proteins are derived from an EF-hand ancestor. *Mol. Biol. Evol.* **13**, 990–998 (1996).
- Lavoie, C., Meerloo, T., Lin, P. & Farquhar, M. G. Calnuc, an EF-hand Ca^{2+} -binding protein, is stored and processed in the golgi and secreted by the constitutive-like pathway in AtT20 cells. *Mol. Endocrinol.* **16**, 2462–2474 (2002).
- Petersson, U. *et al.* Nucleobindin is produced by bone cells and secreted into the osteoid, with a potential role as a modulator of matrix maturation. *Bone* **34**, 949–960 (2004).
- Kanai, Y. *et al.* Natural occurrence of Nuc in the sera of autoimmune-prone MRL/lpr mice. *Biochem. Biophys. Res. Commun.* **196**, 729–736 (1993).
- Jia, D. M. *et al.* Troglitazone prevents and reverses dyslipidemia, insulin secretory defects, and histologic abnormalities in a rat model of naturally occurring obese diabetes. *Metabolism* **49**, 1167–1175 (2000).
- Zhou, A., Webb, G., Zhu, X. & Steiner, D. F. Proteolytic processing in the secretory pathway. *J. Biol. Chem.* **274**, 20745–20748 (1999).
- Duckert, P., Brunak, S. & Blom, N. Prediction of protein convertase cleavage sites. *Protein Eng. Design Select.* **17**, 107–112 (2004).
- Pritchard, L. E., Tumbull, A. V. & White, A. Pro-opiomelanocortin processing in the hypothalamus: impact on melanocortin signaling and obesity. *J. Endocrinol.* **172**, 411–421 (2002).
- Cone, R. D. Anatomy and regulation of the central melanocortin system. *Nature Neurosci.* **8**, 571–578 (2005).
- Phillips, M. S. *et al.* Leptin receptor missense mutation in the fatty Zucker rat. *Nature Genet.* **13**, 18–19 (1996).
- Fan, W. *et al.* Role of melanocortinergic neurons in feeding and the agouti obesity syndrome. *Nature* **385**, 165–168 (1997).
- Mountjoy, K. G. *et al.* Localization of the melanocortin-4 receptor (MC4-R) neuroendocrine and autonomic control circuits in the brain. *Mol. Endocrinol.* **8**, 1298–1308 (1994).
- Liu, H. *et al.* Transgenic mice expressing green fluorescent protein under the control of the melanocortin-4 receptor promoter. *J. Neurosci.* **23**, 7143–7154 (2003).
- Butler, A. A. *et al.* Melanocortin-4 receptor is required for acute homeostatic response to increased dietary fat. *Nature Neurosci.* **4**, 605–611 (2001).
- Taniguchi, N. *et al.* The postmitotic growth suppressor necdin interacts with a calcium-binding protein (NEFA) in neuronal cytoplasm. *J. Biol. Chem.* **275**, 31674–31681 (2000).
- Satoh, T. *et al.* Activation of peroxisome proliferator-activated receptor- γ stimulates the growth arrest and DNA-damage inducible 153 gene in non-small cell lung carcinoma cells. *Oncogene* **21**, 2171–2180 (2002).
- Tsuchiya, T., Shimizu, H., Horie, T. & Mori, M. Expression of leptin receptor in lung: leptin as a growth factor. *Eur. J. Pharmacol.* **365**, 273–279 (1999).
- Oh-I, S. *et al.* Molecular mechanisms associated with leptin resistance: $n - 3$ polyunsaturated fatty acids induce alterations in the tight junction of the brain. *Cell Metab.* **1**, 331–341 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Taguchi, M. Yoshida and E. Kada for technical assistance, and K. Yoshikawa and M. Kojima for supplying mouse NUCB2 cDNA and developing ELISA, respectively. This work was supported in part by grants-in-aid from the Ministry of Health, Labor and Welfare of Japan (to M. Mori).

Author Contributions S.O. and H.S. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M. (mmori@med.gunma-u.ac.jp).

The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates

Ivan Ahel¹, Ulrich Rass¹, Sherif F. El-Khamisy^{2,3}, Sachin Katyal⁴, Paula M. Clements², Peter J. McKinnon⁴, Keith W. Caldecott² & Stephen C. West¹

Ataxia oculomotor apraxia-1 (AOA1) is a neurological disorder caused by mutations in the gene (*APT*X) encoding aprataxin^{1,2}. Aprataxin is a member of the histidine triad (HIT) family of nucleotide hydrolases and transferases³, and inactivating mutations are largely confined to this HIT domain. Aprataxin associates with the DNA repair proteins XRCC1 and XRCC4, which are partners of DNA ligase III and ligase IV, respectively^{4–7}, suggestive of a role in DNA repair. Consistent with this, *APT*X-defective cell lines are sensitive to agents that cause single-strand breaks and exhibit an increased incidence of induced chromosomal aberrations^{4,5,8}. It is not, however, known whether aprataxin has a direct or indirect role in DNA repair, or what the physiological substrate of aprataxin might be. Here we show, using purified aprataxin protein and extracts derived from either *APT*X-defective chicken DT40 cells or *Aptx*^{−/−} mouse primary neural cells, that aprataxin resolves abortive DNA ligation intermediates. Specifically, aprataxin catalyses the nucleophilic release of adenylate groups covalently linked to 5′-phosphate termini at single-strand nicks and gaps, resulting in the production of 5′-phosphate termini that can be efficiently rejoined. These data indicate that neurological disorders associated with *APT*X mutations may be caused by the gradual accumulation of unrepaired DNA strand breaks resulting from abortive DNA ligation events.

The reaction mechanism for mammalian DNA ligases involves three steps: (1) interaction of the ligase with ATP to form an enzyme–adenylate complex; (2) transfer of the activated AMP of the DNA ligase–adenylate complex to the 5′ phosphate at the nick; and (3) formation of the phosphodiester bond and AMP release (Supplementary Fig. 1a)⁹. The second and third steps normally occur by a concerted reaction. Abortive intermediates, however, may be formed when DNA ligases attempt to repair non-ligatable or ‘dirty’ breaks induced by reactive oxygen species, leaving an adenylate group covalently bound to the 5′ phosphate at the nick. The interaction of aprataxin with DNA ligase complexes, and the sensitivity of *APT*X-defective cells to agents that cause single-stranded (ss)DNA breaks, prompted us to test whether the protein might possess an AMP-hydrolase activity that could repair abortive DNA ligation intermediates.

Abortive ligation reactions were carried out to generate an adenylated DNA substrate (Fig. 1a, b). The resulting purified nicked duplex product (DNA V, Fig. 1a) was 36 nucleotides in length and had an AMP group covalently bound to the 5′-terminal phosphate at the site of a nick positioned between two 18-mer oligonucleotides. Human aprataxin, purified as a glutathione S transferase (GST)–His-tagged protein (Supplementary Fig. 2a), where GST and His tags are fused to the amino-terminus of aprataxin (hereafter referred to as

GST–His¹⁰ APTX), acted on this DNA–adenylate complex to catalyse removal of the AMP, as seen by the increased mobility of the adenylated ³²P-labelled 18-mer oligonucleotide during denaturing polyacrylamide gel electrophoresis (PAGE; Fig. 1c, lanes 2–5). Direct confirmation of the removal of AMP was obtained by ³²P-labelling the AMP moiety itself, rather than the DNA (Supplementary Fig. 3,

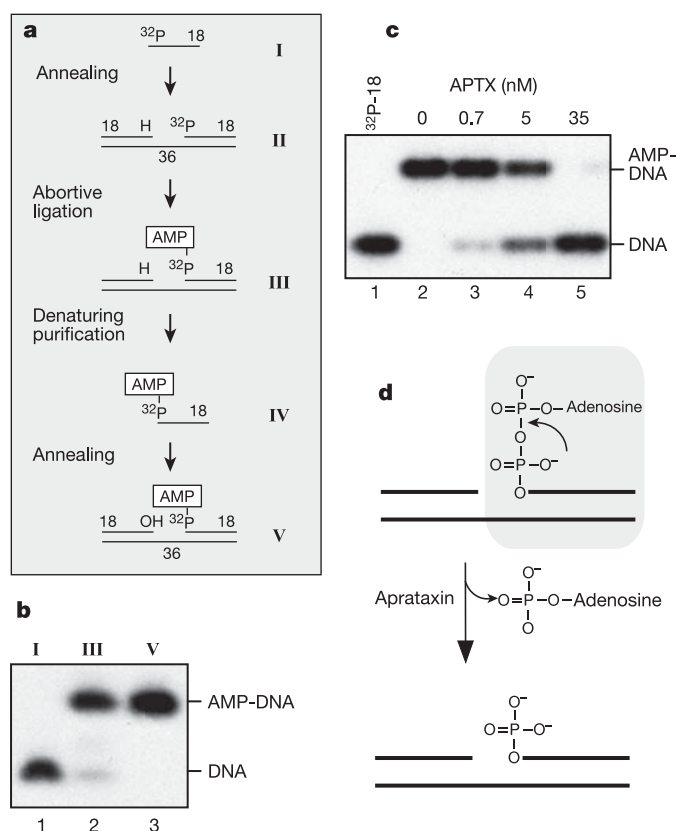


Figure 1 | Removal of AMP from abortive DNA ligation intermediates by aprataxin. **a**, Schematic diagram for the preparation of nicked DNA–adenylates. **b**, ³²P-labelled DNA species (I, III, V) from different stages of the preparation were analysed by denaturing PAGE and autoradiography. **c**, Aprataxin removes AMP from nicked DNA–adenylate. Reactions contained DNA–adenylate (1 μM DNA V) and the indicated amounts of GST–His¹⁰ APTX. Products were analysed by denaturing PAGE and detected by autoradiography. Lane 1, ³²P-labelled 18-mer marker. **d**, Reaction mechanism.

¹Cancer Research UK, London Research Institute, Clare Hall Laboratories, South Mimms, Herts EN6 3LD, UK. ²Genome Damage and Stability Centre, University of Sussex, Falmer, Brighton BN1 9RQ, UK. ³Biochemistry Department, Faculty of Pharmacy, Ain Shams University, PO Box 11566, Cairo, Egypt. ⁴Department of Genetics and Tumor Cell Biology, St. Jude Children's Research Hospital, 332N Lauderdale, Memphis, Tennessee 38105, USA.

lane 3). We found no evidence that aprataxin could act on AMP bound to the active site lysine of DNA ligase (Supplementary Fig. 4). These results show that aprataxin acts on adenylated DNA intermediates to promote pyrophosphate bond hydrolysis and bring about the removal of covalently bound AMP (Fig. 1d).

The above experiments were carried out with a $GST-His$ APTX fusion protein. To eliminate the possibility that the tags (which helped purification) affected aprataxin's activity, they were removed by thrombin treatment (Supplementary Fig. 2b, lane 3). We found that untagged aprataxin exhibited the same activity on the nicked DNA–adenylate as $GST-His$ APTX (Fig. 2a, lanes 3 and 4).

To confirm that aprataxin removes covalently bound AMP to reveal ligatable 5'-phosphate and 3'-OH termini, a ^{32}P -labelled DNA–adenylate was treated with aprataxin in the presence of a variety of DNA ligase–AMP complexes. In the absence of aprataxin, none of the enzyme–adenylate complexes tested (DNA ligase III–XRCC1, DNA ligase IV–XRCC4 or T4 DNA ligase) could ligate the nicked DNA–adenylate substrate (Fig. 2b, lanes 2, 4 and 6). These results show that pre-adenylated ligases fail to process adenylated DNA, presumably because bound AMP blocks their active sites. In the presence of aprataxin, however, efficient nick ligation was seen by the conversion of the adenylated 18-mer into a 36-mer oligonucleotide (Fig. 2b, lanes 3, 5 and 7).

These results indicate that aprataxin removes AMP from the nicked DNA–adenylate to leave a 5' phosphate at the nick that can be ligated to a proximal 3'-OH group (Fig. 2c). Our preliminary observations indicate that the activity of aprataxin on the nicked DNA–adenylate substrate is three to four orders of magnitude greater than that observed previously with AMP–NH₂, or other model substrates such as lysyl-AMP and AppppA¹⁰. The high catalytic activity observed with the nicked duplex DNA–adenylate (the activity is reduced approximately sevenfold with ssDNA–adenylate)

leads us to suggest that this DNA lesion is a prime candidate for being aprataxin's physiological substrate.

To confirm that the DNA hydrolase activity was intrinsic to aprataxin, we made an active site mutant, APTX_{H260A} (mutated at amino acid 260, the central histidine within the H ϕ H ϕ H ϕ motif of the HIT domain, where ϕ represents a hydrophobic residue). Histidine 260 corresponds to the key catalytic residue of known HIT proteins involved in nucleophilic attack on adenylate alpha-phosphate groups³, and mutations of homologous residues in other family members completely abolish their activities¹¹. We found that $GST-His$ APTX_{H260A} failed to promote the hydrolysis of the DNA–adenylate, even at very high enzyme levels (Fig. 3a, lanes 7–10).

Phylogenetic analysis of the HIT protein superfamily indicates that aprataxins from yeast to humans form a discrete branch, suggestive of a common function (Supplementary Fig. 5). To determine whether the specific hydrolytic activity observed with human aprataxin extends to its yeast orthologue, we purified recombinant yeast aprataxin (His-tagged Hnt3) and analysed its activity. We found that Hnt3 also promotes the efficient hydrolysis of the DNA–adenylate complex leading to AMP removal (Fig. 3b, lane 3).

Previous studies have shown that most proteins of the HIT superfamily hydrolyse Ap_nA model substrates³. The HINT1 (histidine triad nucleotide binding) subgroup is known to hydrolyse model substrates in which AMP is joined by a phosphoramidate linkage to the ϵ amino group of lysine^{11–13}, whereas the FHIT (fragile histidine triad)-related proteins hydrolyse dinucleoside polyphosphates such as ApppA and AppppA into mononucleotides^{14–16}. In contrast, proteins in the GalT branch of the HIT superfamily promote the transfer of the nucleotide to a second substrate (that is, they possess transferase activity). We therefore took advantage of yeast strains¹⁷ carrying genomically TAP-tagged versions of HIT superfamily members, to determine whether they share aprataxin's DNA–adenylate hydrolase activity. We found that Hnt1 (HINT1), Hnt2 (FHIT), Apa1 and Apa2 (these latter two proteins have no known orthologues in higher eukaryotes but are mechanistically

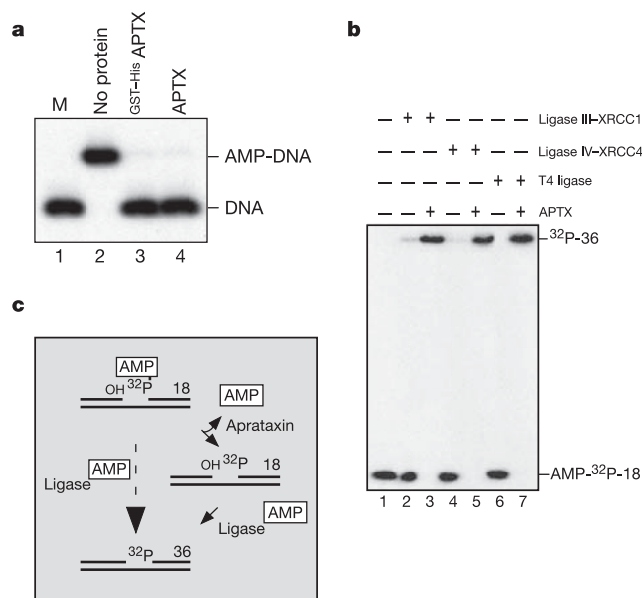


Figure 2 | Aprataxin facilitates the ligation of DNA–adenylate intermediates. **a**, DNA–adenylate hydrolysis catalysed by $GST-His$ APTX or untagged aprataxin (20 nM), as described for Fig. 1c. Lane 1, unadenylated control; lane 2, DNA–adenylate. **b**, Ligation reactions contained the indicated proteins. After 3 min incubation with ATP to generate enzyme–AMP complexes, ^{32}P -labelled DNA–adenylate was added and incubation continued. Products were analysed by denaturing PAGE. **c**, Schematic diagram for the reactions of **b**. Aprataxin catalyses AMP release from the DNA–adenylate by pyrophosphate bond hydrolysis. The 3'-OH and 5'-phosphate termini at the nick may then be ligated by ligase–adenylate complex. The dashed arrow indicates a reaction that cannot be catalysed by ligase–adenylate complex.

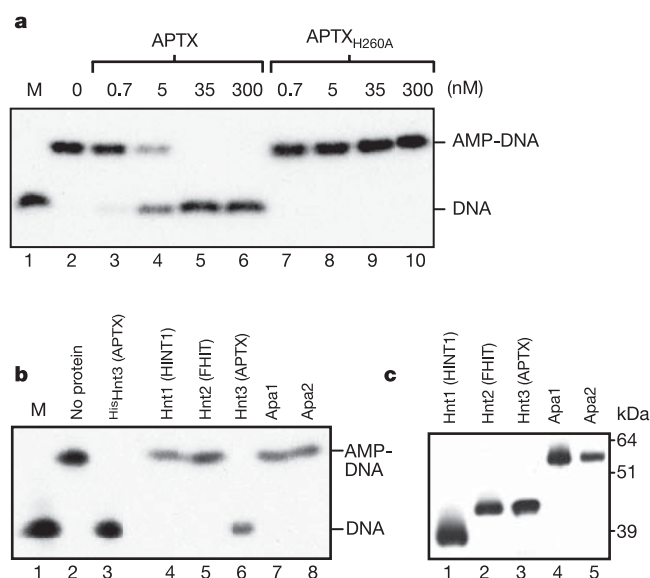


Figure 3 | DNA–adenylate hydrolysis activity is unique to aprataxin. **a**, Active site mutation of $GST-His$ APTX ($GST-His$ APTX_{H260A}) abolishes the hydrolytic activity on adenylated DNA. **b**, Analysis of DNA–adenylate hydrolase activity of the HIT family from yeast. Reactions were carried out using ^{32}P -labelled DNA–adenylate using recombinant His-tagged Hnt3 (35 nM, lane 3) or equal amounts of IgG immunoprecipitates of yeast TAP-tagged proteins (Hnt1, Hnt2, Hnt3, Apa1 and Apa2). Lane 1, unadenylated control; lane 2, mock-treated DNA–adenylate. **c**, Western blot analysis of TAP-tagged yeast HIT family proteins.

with axonal neuropathy-1 (SCAN1)^{24,30}. We therefore suggest that the molecular defect associated with AOA1 relates to the accumulation of unrepaired DNA-adenylates, a direct consequence of the loss of aprataxin activity.

METHODS

See Supplementary Information for details of purified proteins and extracts, DNA substrates, ligation reactions, and aprataxin-targeted mice.

DNA-adenylate hydrolysis by aprataxin. Reactions (5 µl) contained ³²P-labelled nicked duplex DNA-adenylate (1 µM), recombinant aprataxin, immunoprecipitates or whole-cell extracts in hydrolysis reaction buffer (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 5 mM dithiothreitol). Incubation was for 2 min at room temperature (human proteins), 5 min at 30 °C (yeast proteins), or 15 min at room temperature (whole-cell extracts). Reactions were stopped and denatured DNA was analysed by 10% denaturing PAGE followed by autoradiography.

Repair of a 5'-adenylated oxidative single-strand breaks. Mouse cortical astrocyte extracts (8 µg total protein) were pre-incubated at 37 °C for 2 min, followed by incubation for 30 min at 37 °C with 200 nM ³²P-labelled gapped duplex substrate in 25 mM HEPES, pH 8.0, 130 mM KCl, 1 mM dithiothreitol, 10 mM MgCl₂, 1 mM ATP and 100 µM dNTPs. Complementation assays were supplemented with recombinant human His-tagged aprataxin (72 nM). Reactions were stopped by addition of formamide loading buffer and the products analysed by 11% denaturing PAGE and autoradiography.

Received 7 June; accepted 17 August 2006.

Published online 10 September 2006.

1. Date, H. *et al.* Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. *Nature Genet.* **29**, 184–188 (2001).
2. Moreira, M. C. *et al.* The gene mutated in ataxia-ocular apraxia 1 encodes the new HIT/Zn-finger protein aprataxin. *Nature Genet.* **29**, 189–193 (2001).
3. Brenner, C. Hint, Fhit, and GalT: Function, structure, evolution and mechanism of three branches of the histidine triad superfamily of nucleotide hydrolases and transferases. *Biochemistry* **41**, 9003–9014 (2002).
4. Clements, P. M. *et al.* The ataxia-oculomotor apraxia 1 gene product has a role distinct from ATM and interacts with the DNA strand break repair proteins XRCC1 and XRCC4. *DNA Repair (Amst.)* **3**, 1493–1502 (2004).
5. Gueven, N. *et al.* Aprataxin, a novel protein that protects against genotoxic stress. *Hum. Mol. Genet.* **13**, 1081–1093 (2004).
6. Sano, Y. *et al.* Aprataxin, the causative protein for EAOH is a nuclear protein with a potential role as a DNA repair protein. *Ann. Neurol.* **55**, 241–249 (2004).
7. Luo, H. *et al.* A new XRCC1-containing complex and its role in cellular survival of methyl methanesulfonate treatment. *Mol. Cell. Biol.* **24**, 8356–8365 (2004).
8. Mosesso, P. *et al.* The novel human gene aprataxin is directly involved in DNA single-strand break repair. *Cell. Mol. Life Sci.* **62**, 485–491 (2005).
9. Tomkinson, A. E., Vijayakumar, S., Pascal, J. M. & Ellenberger, T. DNA ligases: structure, reaction mechanism and function. *Chem. Rev.* **106**, 687–699 (2006).
10. Kijas, A. W., Harris, J. L., Harris, J. M. & Lavin, M. F. Aprataxin forms a discrete branch in the HIT (histidine triad) superfamily of proteins with both DNA/RNA binding and nucleotide hydrolase activities. *J. Biol. Chem.* **281**, 13939–13948 (2006).
11. Bieganski, P. *et al.* Adenosine monophosphoramidase activity of Hint and Hnt1 supports function of Kin28, Ccl1, and Tfb3. *J. Biol. Chem.* **277**, 10852–10860 (2002).
12. Brenner, C. *et al.* Crystal structures of Hint demonstrate that histidine triad proteins are GalT-related nucleotide-binding proteins. *Nature Struct. Biol.* **4**, 231–238 (1997).
13. Krakowiak, A. *et al.* Biochemical, crystallographic, and mutagenic characterization of Hint, the AMP-lysine hydrolase, with novel substrates and inhibitors. *J. Biol. Chem.* **279**, 18711–18716 (2004).
14. Barnes, L. D. *et al.* Fhit, a putative tumor suppressor in humans, is a dinucleoside 5',5''-P₁P₃-triphosphate hydrolase. *Biochemistry* **35**, 11529–11535 (1996).
15. Lima, C. D., Klein, M. G. & Hendrickson, W. A. Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science* **278**, 286–290 (1997).
16. Draganescu, A., Hodawadekar, S. C., Gee, K. R. & Brenner, C. Fhit-nucleotide specificity probed with novel fluorescent and fluorogenic substrates. *J. Biol. Chem.* **275**, 4555–4560 (2000).
17. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
18. Plateau, P., Fromant, M., Schmitter, J. M., Buhler, J. M. & Blanquet, S. Isolation, characterization, and inactivation of the Apa1 gene encoding yeast diadenosine 5',5''-P₁P₄-tetrakisphosphate phosphorylase. *J. Bacteriol.* **171**, 6437–6445 (1989).
19. Henle, E. S. & Linn, S. Formation, prevention, and repair of DNA damage by iron/hydrogen peroxide. *J. Biol. Chem.* **272**, 19095–19098 (1997).
20. Pogozelski, W. K. & Tullius, T. D. Oxidative strand scission of nucleic acids: routes initiated by hydrogen abstraction from the sugar moiety. *Chem. Rev.* **98**, 1089–1108 (1998).
21. Ward, J. F. in *DNA Damage and Repair* Vol. 2 (eds Nicoloff, J. A. & Hoekstra, M. F.) 65–85 (Humana, Totowa, New Jersey, 1998).
22. Loizou, J. I. *et al.* The protein kinase CK2 facilitates repair of chromosomal DNA single-strand breaks. *Cell* **117**, 17–28 (2004).
23. Whitehouse, C. J. *et al.* XRCC1 stimulates human polynucleotide kinase activity at damaged DNA termini and accelerates DNA single-strand break repair. *Cell* **104**, 107–117 (2001).
24. Caldecott, K. W. DNA single-strand break repair and spinocerebellar ataxia. *Cell* **112**, 7–10 (2003).
25. Seidle, H. F., Bieganski, P. & Brenner, C. Disease-associated mutations inactivate AMP-lysine hydrolase activity of Aprataxin. *J. Biol. Chem.* **280**, 20927–20931 (2005).
26. Barnes, D. E. DNA damage: Air-breaks? *Curr. Biol.* **12**, R262–R264 (2002).
27. Barnes, D. E., Stamp, G., Rosewell, I., Denzel, A. & Lindahl, T. Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Curr. Biol.* **8**, 1395–1398 (1998).
28. Gao, Y. *et al.* A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* **95**, 891–902 (1998).
29. Abner, C. W. & McKinnon, P. J. The DNA double-strand break response in the nervous system. *DNA Repair (Amst.)* **3**, 1141–1147 (2004).
30. El Khamisy, S. F. *et al.* Defective DNA single-strand break repair in spinocerebellar ataxia with axonal neuropathy-1. *Nature* **434**, 108–113 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Barnes and T. Lindahl for discussions and the gift of DNA ligase III-XRCC1 complex, and A. Ciccia for advice and comments. We thank S. Rulten for the gift of recombinant His-tagged aprataxin, and M. Taylor for provision of AOA1 lymphoblastoid cells. This work was supported by Cancer Research UK (S.C.W.), the EU DNA Repair Consortium (S.C.W.), the Medical Research Council (K.W.C.) and the NIH (P.J.M.). I.A. is supported by a fellowship from the European Molecular Biology Organisation.

Author Contributions I.A. made the initial discovery of DNA-adenylate activity, and I.A. and U.R. were responsible for experimental design and the generation of most of the experimental data. S.F.E.-K., S.K. and P.M.C. were responsible for the development of vertebrate models and performed some of the experiments. P.J.M. and K.W.C. provided critical resources and expertise, and contributed to writing the manuscript. S.C.W. was the project leader and produced the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.C.W. (stephen.west@cancer.org.uk).

ADDENDUM

doi:10.1038/nature05188

Sustainability of three apple production systemsJohn P. Reganold, Jerry D. Glover, Preston K. Andrews
& Herbert R. Hinman*Nature* 410, 926–930 (2001)

We wish to clarify links between us and Stemilt Growers Inc., a commercial fruit company in Washington State, USA, regarding this Letter. The study site was part of a commercial orchard in the Yakima Valley, Washington. At the start of the study in 1994, Stemilt Growers Inc. did not own any part of the farm. Farmer Andy Dolph was a field consultant for Stemilt Growers, but the company had no involvement with the design, location or assessment of the experiment. In 1996 Stemilt Growers bought part of the farm, but they took over management of the farm only at the end of the 2001 growing season, whereas our *Nature* paper covered previous growing seasons. Stemilt Growers created the Responsible Choice environmental-impact rating system used in our study but had nothing to do with our decision to use this rating system, did not suggest that we should examine environmental impact, and were not involved in our evaluation of the farming systems. A more detailed version of this statement is available online as Supplementary Information to this Addendum.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

CORRIGENDUM

doi:10.1038/nature05086

Structure of the *E. coli* protein-conducting channel bound to a translating ribosome

Kakoli Mitra, Christiane Schaffitzel, Tanvir Shaikh, Florence Tama, Simon Jenni, Charles L. Brooks, Nenad Ban & Joachim Frank

Nature 438, 318–324 (2005)

We wish to withdraw a poorly worded criticism of Arthur E. Johnson's fluorescent studies of the ribosome–PCC complex, contained in the Supplementary Information to this Article. The issues raised by the discrepancies between several cryo-EM studies and the body of work from Johnson's group regarding solvent accessibility of the ribosome–PCC junction warrant more extensive study and discussion in the scientific community. The withdrawal of the statement has no bearing on the results of our study or on their interpretation.

Picasso's cat

The editor's wavefunction collapses.

Ron Collins

Dear Mr Gee,

Thank you for your recent correspondence in regard to an earlier submission. Some of your comments have gotten me to thinking, and as a result I now understand a few things more fully than I once did. Knowing your fascination with topics as eclectic as quantum theory, art history and, of course, editorial matters, I immediately thought you would want to hear about my recent endeavours, and how I've tied these apparently disparate fields together into a nice little ball.

What I've found is certainly most disturbing, but my results can be verified through simple logic.

In other words, I'm on to you.

It all started, I now understand, with the cubists.

You know, Picasso and the rest of the crowd that took pieces of people and cows and dogs and for all I know anything else and chopped them up, placing a leg here, an arm there, and a breast at the eye level of the average 35-year-old man.

After several hours of research, I am convinced — and I'm certain you are, too — that the cubists knew well something it took Bohr, Heisenberg and Schrödinger another 25 or so years to almost figure out.

That is, of course, the cubists knew a good deal about quantum theory.

I refer to the example of Schrödinger's cat. In this famous thought-experiment a cat is placed in a box and predictions are made as to whether it survives or not. Quantum theorists then proceed to state that until the box is opened, the cat is in the astonishing state of fluctuation between both possibilities — alive half the time and dead the other half (which must be terribly unpleasant for the cat). I'm an engineer, though, not a theoretical physicist, and whether the cat is in oscillation between life and death or merely just sleeping is a philosophical discussion at best. So I'll leave that for our next communication in order to concentrate on hard fact.

So, let's turn back to the world of art. The proof, they say, is in the pudding, and simply by looking at their work, I think it an obvious deduction that the cubists had discovered quantum theory and were experimentally proving Schrödinger's dalianace well before he thought it all out. Only they went a step farther than Schrödinger,

applying the theorem recursively to various parts of the anatomy of their subjects.

To visualize what I mean, think of Schrödinger's cat, but then assign each portion of the cat its own quantum box — one for the left eye, one for the nose, a few for long whiskers, one for the tail. You get the point. Once you get this image, it's simple to understand how the artist could render such visions as the cubists did, for their reality would depend only upon which of the multitudes of boxes they chose to look into and what state each element lay in when it was observed.



The fact that the cubist movement even exists is obvious proof that the uncertainty principle works on the large scale as well as at the subatomic and, in fact, has a recursive principle. Luckily, however, we live in a moderately stable universe where very few people understand how to manipulate the quantum flux like the cubists did. Otherwise, I figure we would all look like Picasso's cat.

Now that I understand it, I'm really quite happy to know that you are one of the very few to be actively employing this principle.

I told you I was on to you.

I hear you, Henry, I really do: how did

I, a simple engineer with limited faculties, figure it all out where so many other really bright folks have failed?

Well, it goes like this.

I was looking in my e-mail the other day, and I thought, as usual, "Is there a message from Henry inside my inbox, or is there no message from Henry inside the inbox?" Schrödinger's post, as it were. Suddenly, it hit me. I've been asking the wrong question.

Instead, I should be asking recursive, Picasso-post questions such as "Assuming there is a message inside, will it be signed Henry, or will it be signed Mr Gee?"; "Will it be type-written or will it be a graphic insert?"; and "Will the salutation be 'Dear Ron' or 'Dear Mr Collins?'" You know — identifying the source of the components of the message, as the cubists did, rather than applying the theory to the whole of the cat, as the physicists did.

As I processed this thought, I suddenly understood what you were doing.

All those rejection letters are merely your way of ferreting out less intelligent authors and general troublemakers in favour of those with such moxie as yours. You've been sending me Picasso's memoranda, quantum correspondence, letters in flux between acceptance and rejection. And while I've enjoyed your commentary, I've rarely looked close enough to open the right reality, the one that resulted in the elusive cheque rather than the dreaded letter of declination.

So I wanted to tell you that I was on to you.

I can't go back, of course. Once an envelope is opened, the rejection cannot be changed. So most of the tales I've sent you so far will certainly have to find resting homes with some other, inferior magazine.

But I promise I'll await your next letter with eager anticipation.

Regards,
Ron Collins

P.S. I've been trying this theory out in preparation for your reply, and I think I've got it about right.

P.P.S. You should see my cat.

Ron Collins has contributed nearly 40 stories to professional publications, including *Analog*, *Dragon* and several original anthologies. Later this year, his story '1 is True' will appear in *Asimov's*.

JACEY